Learned Navigation of StyleGAN3 Latent Space from Audio Descriptors

Ted Moore
Computer Music Department
Peabody Institute of Johns Hopkins University
Baltimore, MD USA
tmoore97@jh.edu

Abstract

This paper introduces a machine learning workflow that emphasizes having an artist-in-the-loop for creating audio-reactive videos by navigating the latent space of StyleGAN3 using audio descriptor analyses. The artist pairs high dimensional audio analyses conducted with the FluCoMa Toolkit with latent vectors of a custom-trained StyleGAN3 model using a flexible two dimensional plot and real-time audio scrubbing. A neural network is trained to predict the latent vectors from the paired audio analyses. By prioritizing small, artist-curated datasets and using artist-in-the-loop supervised learning processes, this system enables artists to embed personally meaningful and abstract aesthetic decisions into a machine learning workflow. This approach avoids the limitations of many existing methods, such as a reliance on semantic audio-visual mappings and large, impersonal datasets. Rather than focusing technical novelty, this paper proposes a methodology that centers an artist-in-the-loop workflow integrating subjective artistic criteria with machine learning algorithms. A case study for violin, tape, and video, demonstrates the system's ability to create tight integration where the correlations between audio and visual elements are immediate and specific.

1 Introduction

This study presents a machine learning workflow to create audio-reactive videos by navigating StyleGAN3's (Karras et al., 2021) image-generating latent space using audio descriptor analyses. Images generated from StyleGAN3's latent vectors are manually paired with audio descriptors creating a completely artist-defined and manageably-sized, multi-modal dataset enabling a dynamic relationship between audio and video media. This system improves previous research by using audio analyses capable of differentiating widely varying musical moments including noisy and un-pitched sounds. Further, the results are more audio-reactive than previous studies on a moment-to-moment basis, more explicitly responding to short-term sonic morphologies.

The case study for this system was a composition for violin, video, and tape, titled *angle*, in which the author wanted to synchronize the video-morphing potential of StyleGAN3 with the tape part. Audio-visual results of the proposed workflow can be seen here.

1.1 Previous Work

The StyleGAN family of model architectures (Karras, 2019; Karras et al., 2020, 2021) is well known for the high quality images they generate. The quality of images has led some researchers to use StyleGAN and other Generative Adversarial Networks (GANs) (Goodfellow et al., 2020) to generate images of spectrograms for audio resynthesis. (Palkama et al., 2020) In one case, spectrograms of foley effects were inferred from the semantic content of video frames. (Ghose and Prevost, 2022) Because of GANs common association with face generation, previous research that generating video by navigating StyleGAN's latent space via audio analyses has often aimed at matching moving lips

Proceedings of the 6th Conference on AI Music Creativity (AIMC 2025),
Brussels, Belgium, September 10th-12th

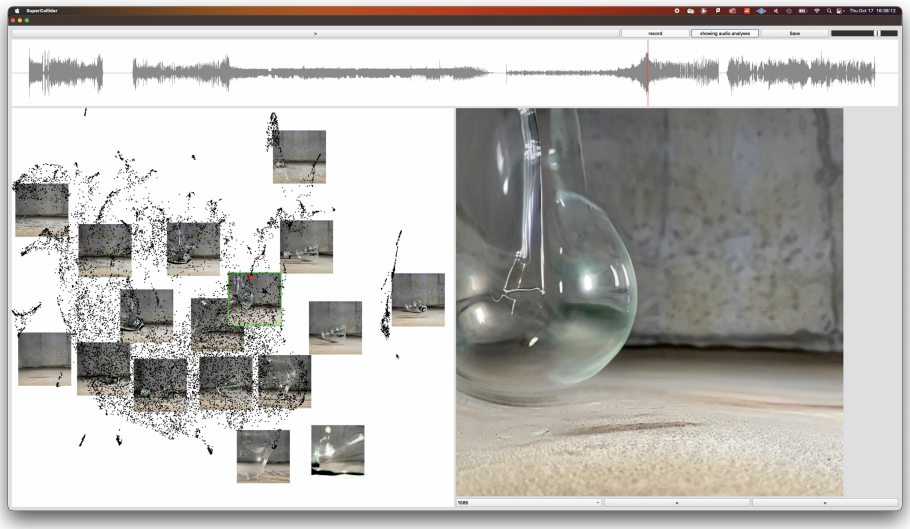

Figure 1: Screen capture of the two dimensional plot for pairing latent vector-generated images with audio analyses.

to a recorded voice, (Tan et al., 2024; Yin et al., 2022; Su et al., 2024) some including completely synthesized video. One study which even aimed to generate a synthetic video of a speaking face presenting a gender inferred from a target recorded voice. (Ngo) Contrastive learning models creating multi-modal embeddings that include sound (Chen et al., 2022; Wu* et al., 2023; Ilharco et al., 2021) have enabled *semantic* generation and manipulation of images (Lee et al., 2022) and video such as generating video of fire from sounds of fire using StableDiffusion as the video-frame generator. (Jeong et al., 2023)

Because all of the above studies focus on pairing sounds and visuals based on semantic meaning, they are not suited for abstract artistic expression. While (Jeong et al., 2021) creates StyleGAN-generated, audio-reactive videos by pairing audio and visual elements using subjective aesthetic criteria, their results are limited by the choice of audio analyses used (onsets and chromagram) and an overuse of smoothing, leading to a reliance on pitch- and chord- based music and a lack of immediacy in the audio-reactivity. Additionally, because many of the preceding studies, especially those using contrastive learning, aim to create a model generalized to a variety of real world applications, they require datasets larger than are feasible or meaningful for individual artists to create, manage, and be expressive with.

This paper proposes an aesthetic ideology different from the studies above: a workflow for individual artists who want to use manageably-sized, personally-created datasets that have specific meaning and content. (Rachel Fensham, 2023; Wang, 2019) While similar artist-in-the-loop approaches exist, such as Akten's *Ultrachunk* (2018) using mel-spectrograms, (Dyer et al., 2023) this workflow differs by: (1) using richer audio descriptors that capture more timbral and spectral characteristics, (2) providing a more flexible interface for abstract audio-visual pairing, and (3) leveraging StyleGAN3's improved latent space organization for better interpolation quality enabling coherent visual morphologies abstracted from the source dataset.

This workflow creates a model that expresses an audio-visual relationship intended to exist inside one composition or creative project. The same system can be used with different source images and sounds creating new models and different artistic expressions for different projects. The strategies in this study also model how artists can use machine learning creatively while actively avoiding ethical conflicts that can arise from using certain data, such as copyright infringement, privacy concerns, and labor exploitation. (Crawford and Paglen, 2021; D'ignazio and Klein, 2023)

2 Image Dataset & StyleGAN3 Training

The visual material for training StyleGAN3 was created by recording videos[1] of glass objects (bottles and light bulbs) falling and breaking. These videos were edited down to specific moments that were aesthetically desirable and then exported to a `png` sequence, center-cropped to squares, and resized to `1024x1024` pixels for training with StyleGAN3, resulting in 44,073 images. Training began from a Flickr-Faces-HQ (FFHQ) checkpoint[2] (Karras, 2019) and trained for 480,000 images.

3 Audio Dataset

In order to represent and differentiate moments in the tape part, every FFT frame of the audio (sample rate = 44100, window size = 1024, hop size = 512) is analyzed in SuperCollider (McCartney, 2002) using 23 descriptors from the FluCoMa Toolkit (Tremblay et al., 2021): spectral centroid, spectral spread, spectral skewness, spectral kurtosis, spectral rolloff, spectral flatness, spectral crest, 13 MFCCs (skipping coefficient 0, starting at coefficient 1), pitch, pitch confidence, and loudness.[3] Any data points with a loudness below -60 dBFS was removed from this dataset. The remaining dataset was standardized to mean of 0 and standard deviation of 1. Using Principal Component Analysis, the top 19 principal components were kept, maintaining greater than 99% of the variance. Finally, Uniform Manifold Approximation and Projection (UMAP) (McInnes et al., 2018) was used to embed[4] the data into a two dimensional space for use as described in Section 4.

4 Pairing Images with Audio

The desired behavior is to have specific audio analyses generate specific images via StyleGAN3 and interstitial audio analyses generate interpolated, interstitial images, therefore creating video morphology tightly corresponding to the sonic morphology. Initial experiments attempted to create the audio-analysis-to-image-generation relationship automatically using various unsupervised techniques including using KMeans clustering, linear assignment, and principal component-matching, however, inserting the artist-into-the-loop by manually pairing the data for supervised learning led to the best results.[5] Experimentation determined that carefully selecting and pairing the audio and image examples had the greatest impact on the visual results, so special care was taken to create a system for managing and paring these data points in a fluid and meaningful way.

First, in order to get a sense of what kinds of visual outputs the trained StyleGAN3 model would provide, 500 random latent vectors (in 512 dimensional space) and their resulting images were generated (for an example, see Figure 2). From these images, 17 were selected in order to represent the variety in the results and for their high image quality since these images will likely end up in the resulting video.

Next, the two dimensional UMAP embedding of the audio analyses was plotted (in SuperCollider using FluidPlotter, see Figure 1) with a visual interface allowing for scrubbing across the sound file, hearing the results, and seeing where on the two dimensional plot each sonic moment was embedded.[6] Overlaid on the plot were the 17 chosen images. These images were manually moved to the two dimensional positions on the plot corresponding to the sound analyses desired to predict the given image.

In this visual interface, when the audio is playing, a 30 frames-per-second mockup is created by displaying the image nearest to the currently heard sound analysis (in the 2 dimensional space). While this "video" of course does not include the StyleGAN3 interpolations it is useful to get a sense of what kinds of target images might be seen in the final video.

[1] 1920x1080 resolution at 120 frames-per-second

[2] `stylegan3-r-ffhqu-1024x1024.pkl`

[3] All FFT analyses ranged from 20 Hz to 10 kHz.

[4] The UMAP parameters were number of neighbors = 15 and minimum distance = 0.1.

[5] as is often the case

[6] Silent moments (below -60 dBFS) initially removed from the dataset were represented by the audio analysis of the most recent non-silent moment

Figure 2: One latent vector-generated image of a light bulb shattering.

Next, each of the 17 target latent vectors need to be paired with one audio analysis vector (creating an admittedly but also desirably *small*-artist-defined dataset). While the visual interface enabling the pairing is in two dimensions, we want to maintain the higher dimensional structures (512 dimensions of the StyleGAN3 latent space and 23 dimensions of the audio analyses) for actually learning these relationships. For each image on the plot, the nearest neighbor audio analysis in the two dimensional space was found. That audio analysis' standardized 23 dimensional vector representation was paired with the image's 512 dimensional latent vector creating a dataset of 17 input-output pairs.

5 Training & Inference

A fully connected neural network with one hidden layer of 40 neurons using the ReLU activation function (and no activation function on the output) was trained (using PyTorch (Paszke et al., 2017)). Mean squared error was used as the loss function with an Adam optimizer beginning at a learning rate of 0.01. Training showed improvement over 500 epochs clearly over-fitting to a loss five orders of magnitude smaller than at the start. Over-fitting in this case is acceptable or even desirable because there is no real-world use-case it needs to generalize to, instead the precision of predictions will ensure the desired images will be generated by StyleGAN3.

In order to create the timeseries of StyleGAN3 latent vectors from the timeseries of standardized audio analyses three things needed to happen: (1) the timeseries needed to be resampled from about 86 Hz (because of the hop size of 512) to 30 Hz (to create a video at 30 frames per second), (2) the latent vectors needed to be inferred from the audio analysis vectors, and (3) smoothing must be applied to the timeseries to smooth out jitter in the audio analyses. A grid search was conducted attempting

different ordering of operations as well as different parameters for smoothing and resampling. After subjective aesthetic assessment of the results, the following process was determined preferable: First, smoothing is applied to the timeseries of standardized audio analyses using a leaky integrator with an coefficient of 0.8 (always between 0-1, higher = more smoothing). Second, the smoothed timeseries of audio analyses is down-sampled to 30 frames-per-second by using the mean value of a moving kernel 1/30th of a second in width. Lastly, the data points in the smoothed and resampled timeseries of audio analyses are used to infer StyleGAN3 latent vectors using the trained neural network. The resulting timeseries of inferred latent vectors are then used for image generation and sequenced into a video using ffmpeg. (FFmpeg Developers, 2024) The source audio was added later in video editing software.

6 Case Study: *angle*

In order to increase the stakes of the experimentation outlined in this study and provide some context for qualitative aesthetic assessment of the results, a piece was composed with the intention of incorporating this workflow. The aesthetic objective is to create an abstract audio-visual composition where visual transformations directly correspond to sonic morphologies so the audience can perceive subtle audio characteristics through their visual representations. This differs from the semantic approaches described above (Ghose and Prevost, 2022; Tan et al., 2024; Yin et al., 2022; Su et al., 2024; Ngo; Chen et al., 2022; Wu* et al., 2023; Ilharco et al., 2021; Lee et al., 2022; Jeong et al., 2023) that rely on conceptual audio-visual associations. Videos of glass objects (bottles and light bulbs) falling and breaking were used to generate the images for training StyleGAN3 (described in Section 2). The tape part (the audio described in Section 3) was constructed of modular synthesizer recordings. The violin part was composed (prior to creating the video as described in this paper) by transcribing the timbre, pitches, dynamics, etc., of the tape part for the violinist to perform.

The author determined the video results of the workflow described above to be compelling in some moments and too noisy/chaotic in other moments. Reflecting on the dataset chosen, noisy sounds and images of shattering glass both have higher entropy than other possible sources, which is more challenging for audio analyses and convolutional neural networks represent in compact dimensions. The author's artistic interest in high entropy media is an interesting challenge to navigate when using algorithms that are more "successful" with lower entropy data. In the resulting video, when the audio-visual coordination was strong, there was a sense of the video being directly controlled by the sonic morphology. In contrast to (Jeong et al., 2021) the author chose to design the system to prioritize frame-to-frame audio-visual correspondence over temporal smoothness. While this creates some abrupt transitions, it ensures immediate responsiveness to sonic changes. The visual "chaos" in certain moments can be understood as an authentic representation of the sonic complexity, suggesting that the system's apparent "failures" may actually constitute a successful translation of audio entropy into visual entropy. Future work can explore temporal modeling approaches that maintain audio-reactivity while improving visual continuity, such as incorporating optical flow constraints, more/different temporal smoothing in the latent space, or lower entropy datasets.

In order to compose these results into *angle*, passages that were particularly compelling were used, while for other moments of the video design, non-StyleGAN3 generated videos of the breaking glass were used to maintain aesthetic cohesion and develop visual motives. Artistic editing of algorithmic outputs is an important part of the author's creative practice, reflecting the belief that subjective aesthetic assessment and human intuition are key elements in creative practice. A completed version of the work can be seen here.

7 Future Research

One challenge with the current workflow is the common problem of the differentiation in the StyleGAN3 latent space. Conditioning the latent space with labels included in the initial training of might make the latent space more differentiable (Brock, 2018) and provide clearer and more interesting interstitial interpolations.

While over-fitting in this case is acceptable because of the very specific use case and the desire for precise predictions, a less over-fit network might create smoother interstitial interpolations because of fewer hyperplanes dividing the high dimensional space creating more linear interpolations. (Prince,

2023) Many strategies such as smaller network sizes, different activation functions, and regularization should be tried. Additionally, using a small dataset such as this, direct regression from audio or audio analyses to images should be explored (skipping over the latent space matching), which might create a stronger sense of audio reactivity at the expense of the GAN's high image quality and well organized latent space.

Lastly, the very piece-meal structure of this system allows for many experiments that may improve performance. One could change the audio analyses, the image-source content, the data preprocessing, the size of the supervised learning dataset, the training process, the smoothing and down-sampling strategies, and more. Each of these is sure to have downstream effects on the results. I hope that others will pick up this architecture and find decisions to tailor it to their idiosyncratic artistic goals.

8 Ethics Statement

All of the data used in training was created by the author. The Flickr-Faces-HQ (FFHQ) dataset has a privacy policy that can be found in their GitHub repo.[7] The author has made efforts to make the research and artistic use of these findings as accessible and inclusive as possible. All of the code is open source and available on GitHub under the GNU General Public License v3 in this repo. The ethical advantages of this approach include: (1) avoiding copyright infringement by using only self-created content, (2) eliminating privacy concerns associated with using others' images or voices, and (3) preventing exploitation of unpaid labor in dataset creation. This methodology can be generalized to other artistic applications where artists wish to maintain ethical control over their machine learning workflows.

All of the StyleGAN3 training and inference was done on a GPU that is part of the Yale University High Performance Computing Grace Cluster. The remaining technical work was performed on a MacBook Pro laptop. All of the SuperCollider and Python code that does the analysis and machine learning has been created with the aim of computational efficiency for minimal energy usage.

9 Conclusions

This paper presents an artist-in-the-loop approach to machine learning-based audio-visual creative practice, emphasizing the importance of a small, personally created dataset and human intuition in the design of a reactive media system. Rather than pursuing generalizable trainings or semantic mappings, this methodology foregrounds subjective intuition and aesthetic specificity, enabling artists to craft personalized abstract correspondences between sound and image. This work contributes to a growing discourse on artist-in-the-loop machine learning by offering a transparent, modifiable methodology for others to adapt to their own creative practices.

Acknowledgments and Disclosure of Funding

Special thanks to the Yale Center for Research Computing, specifically Tom Langford, Jay Kubeck, and Kaylea Nelson, for guidance and assistance in computation run on the Grace cluster.

References

Andrew Brock. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018.

Ke Chen, Xingjian Du, Bilei Zhu, Zejun Ma, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Hts-at: A hierarchical token-semantic audio transformer for sound classification and detection. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, 2022.

Kate Crawford and Trevor Paglen. Excavating ai: The politics of images in machine learning training sets. *Ai & Society*, 36(4):1105–1116, 2021.

Catherine D'ignazio and Lauren F Klein. *Data feminism*. MIT press, 2023.

[7] https://github.com/NVlabs/ffhq-dataset

Mark Dyer, Zubin Kanga, and Jennifer Walshe. In conversation with jennifer walshe: Performing with intelligent machines. *Contemporary Music Review*, 42(3):391–399, 2023. doi: 10.1080/07494467.2023.2276563. URL https://doi.org/10.1080/07494467.2023.2276563.

FFmpeg Developers. Ffmpeg, 2024. URL https://ffmpeg.org. Version 7.1.

Sanchita Ghose and John J Prevost. Foleygan: Visually guided generative adversarial network-based synchronous sound generation in silent videos. *IEEE Transactions on Multimedia*, 25:4508–4519, 2022.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

Gabriel Ilharco, Mitchell Wortsman, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, July 2021. URL https://doi.org/10.5281/zenodo.5143773.

Dasaem Jeong, Seungheon Doh, and Taegyun Kwon. Träumerai: Dreaming music with stylegan. *arXiv preprint arXiv:2102.04680*, 2(4):10, 2021.

Yujin Jeong, Wonjeong Ryoo, Seunghyun Lee, Dabin Seo, Wonmin Byeon, Sangpil Kim, and Jinkyu Kim. The power of sound (tpos): Audio reactive video generation with stable diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7822–7832, 2023.

Tero Karras. A style-based generator architecture for generative adversarial networks. *arXiv preprint arXiv:1812.04948*, 2019.

Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of StyleGAN. In *Proc. CVPR*, 2020.

Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. In *Proc. NeurIPS*, 2021.

Seung Hyun Lee, Wonseok Roh, Wonmin Byeon, Sang Ho Yoon, Chanyoung Kim, Jinkyu Kim, and Sangpil Kim. Sound-guided semantic image manipulation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3377–3386, 2022.

James McCartney. Rethinking the computer music language: Super collider. *Computer Music Journal*, 26(4):61–68, 2002.

Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

Jerry Ngo. Stylewav: Guiding image synthesis using audio.

Kasperi Palkama, Lauri Juvela, and Alexander Ilin. Conditional spoken digit generation with stylegan, 2020. URL https://arxiv.org/abs/2004.13764.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. In *NIPS-W*, 2017.

Simon JD Prince. *Understanding deep learning*. MIT press, 2023.

Signe Ravn Ashley Barnwell Danny Butt Rachel Fensham, Tyne Daile Sumner, editor. *Small Data is Beautiful*. Grattan Street Press, Univeristy of Melbourne, Australia, 2023.

Jiacheng Su, Kunhong Liu, Liyan Chen, Junfeng Yao, Qingsong Liu, and Dongdong Lv. Audio-driven high-resolution seamless talking head video editing via stylegan. In *2024 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6, 2024. doi: 10.1109/ICME57554.2024.10688257.

Shuai Tan, Bin Ji, and Ye Pan. Style2talker: High-resolution talking head generation with emotion style and art style. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 5079–5087, 2024.

Pierre Alexandre Tremblay, Gerard Roma, and Owen Green. Enabling Programmatic Data Mining as Musicking: The Fluid Corpus Manipulation Toolkit. *Computer Music Journal*, 45(2):9–23, 06 2021. ISSN 0148-9267. doi: 10.1162/comj_a_00600. URL https://doi.org/10.1162/comj_a_00600.

Ge Wang. Humans in the loop: The design of interactive ai systems, 10 2019. URL https://hai.stanford.edu/blog/humans-loop-design-interactive-ai-systems.

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 *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, 2023.

Fei Yin, Yong Zhang, Xiaodong Cun, Mingdeng Cao, Yanbo Fan, Xuan Wang, Qingyan Bai, Baoyuan Wu, Jue Wang, and Yujiu Yang. Styleheat: One-shot high-resolution editable talking face generation via pre-trained stylegan. In *European conference on computer vision*, pages 85–101. Springer, 2022.