# Echoes of Art: Translating Audio into Art

**Ryan Ng**
Simon Fraser University
Surrey, BC, Canada
Ran3@sfu.ca

### Abstract

Visualizing sound is an age-old challenge, one that has evolved alongside technology and artistic innovation. From early waveforms to modern visualizations, the attempts to translate audio into visual forms reflect a deeper desire to bridge the sensory experiences of hearing and sight. *Echoes of Art* continues this through the transformation of sound into stylized visual art using spectrograms and Neural Style Transfer. The project transforms audio into a dynamic visual expression, applying the aesthetics of classical paintings like Vincent van Gogh's *The Starry Night* in creating a new, unique art form. Users directly interact with the process through selecting from five artistic styles, engaging directly with the translation of sound into visual art. The results reveal how audio data, when combined with artistic styles, can evoke an emotional response similar to visual art. The project offers an alternative to understanding sound as visual medium and expands the creative potential for cross-sensory interpretations.

### Keywords

Audio visualization, audio-visual art, spectrograms, Neural Style Transfer (NST), Artificial Intelligence (AI), generative art, cross-sensory art.

## Introduction

*Echoes of* Art explores the creative potential of everyday sounds by transforming them into works of art inspired by popular painting styles. By capturing sounds from the environment, and transforming them into works of art, this project challenges how sound, which is traditionally viewed as an auditory sensation, can be reimagined in the visual sense. The core concept of this project is to merge spectrograms, created from captured sound, with elements of well-known painting styles to create works of art that show the uniqueness of everyday sounds.

Through this fusion of sound and art, this project highlights the potential for sound to be used beyond its normal auditory boundaries and engage with visual art in a meaningful way. By incorporating popular artistic styles, the resulting artwork remains familiar and relatable, allowing viewers to connect with it, despite its unique creation process and departure from traditional methods.

Using Neural Style Transfer (NST), a technique that uses artificial intelligence to apply the visual style of one image to another image while preserving its original content, spectrograms are blended with well-known painting styles like Van Gogh's *The Starry Night* or Munch's *The Scream*. Users can choose from five classical painting styles, creating a personalized, interactive experience.

## Motivations

The concept of *Echoes of Art* emerged from a curiosity about how sound could be used beyond its traditional auditory context. This interest deepened during the Layered LED Screen assignment, where I began exploring AI systems as creative tools. That experience helped solidify my approach and inspired a direction that combined sound, art, and machine learning. My goal was to investigate the artistic potential of everyday sounds by transforming them into visual artworks influenced by iconic styles.

## Background

This project incorporates three key elements: audio spectrograms, Neural Style Transfer (NST), and five classical paintings, to transform sound into unique works of art. Spectrograms, generated using the Short-Time Fourier Transform (STFT), provide the initial audio visualization by capturing the frequency content of sound over time. Audio files given by the users are processed and converted into spectrograms, which serve as the base image for the project, allowing for further transformations.

Neural Style Transfer is then applied to the spectrogram, using deep learning techniques to blend the content, which includes the structural elements of the image such as shapes, objects, and overall layout, with the stylistic elements of a classical artwork. By separating the content and style of an image, NST allows the spectrograms to retain their underlying structure while adopting the visual aesthetic of iconic paintings.

The five painting options provided to the user are Vincent van Gogh's *The Starry Night*, Edvard Munch's *The Scream*, Hokusai's *The Great Wave Off Kanagawa*, Claude Monet's *Impression, Sunrise*, and Salvador Dalí's *The Persistence of Memory*. These paintings were intentionally selected based on their popularity, distinct features, and artistic styles allowing for a diverse range of images to be created.

# Literature Review

The idea of turning sound into images is not a new one, but rather one that has been explored in a variety of ways across different disciplines. In my preliminary research, I found several papers and projects that helped guide me to my final concept, each having a unique approach to visualizing sound. While these works use different tools and techniques, they share a common goal of translating sound into visual form. These examples provided valuable insight into how sound can be reimagined using technology and helped shape the direction of *Echoes of Art*. The following sections outline the key contributions and relevance of each work.

## Images that Sound

Chen et al. (2025), in their paper *Images that Sound: Composing Images and Sounds on a Single Canvas*, present a method for generating spectrograms that resemble real images, which can then be enhanced and converted into sound waves [1]. Their approach uses two separate pre-trained models within the same latent space, where a prompt is given to both a text-to-image model as well as a text-to-spectrogram model, producing an output that functions as both an image and a spectrogram. The resulting image can then be colorized to emphasize its visual aspects or transformed into a waveform using a pre-trained vocoder.

While this method differs from my project, it influenced my decision to use spectrograms as the base image for transformation and inspired the concept of converting sound into an image.

## Sound2Scene

Sound2Scene, a sound-to-image generative model presented by Sung-Bin et al. (2023) in their paper *Sound to Visual Scene Generation by Audio-to-Visual Latent Alignment,* is a model that is composed of three parts: an audio encoder that processes the input sound, a pre-trained image encoder, and an image generator [2]. First, the image encoder is trained using images to extract meaningful visual features. Then, the image generator is trained to recreate images using the visual features extracted from the encoder. Once the image generator reaches a high accuracy of generating correct images, the model moves on to learning how to align audio features with visual features by training on videos that have both sound and images. The model then maps audio features into a shared audio-visual representation space, to ensure that sounds corresponding to specific visuals are positioned close together. Once the audio features are properly aligned, the model can start generating images from sound. It takes a sound input, extract its features, and uses the image generator to create a corresponding image.

This approach helped me discover Neural Style Transfer (NST), as I wanted to create images that preserved the entire structure of sound rather than breaking it apart, as Sound2Scene does. I found that focusing on isolated parts of sound didn't fully capture its essence, which is a crucial aspect of my project.

## Soundscape-to-Image

Soundscape-to-Image is a computational framework presented by Zhuang et al. (2024), that transforms audio soundscapes into visual images of places in a street-view format [3]. The Soundscape-to-Image model is trained using a dataset of collected street videos containing both visual and auditory elements. Once trained, the model can generate images of places solely from audio inputs. To ensure the accuracy of the model, four validation tasks are performed to evaluate the reliability and performance of the model.

This model is the closest to what I have created, but the major difference is that it is trained to associate specific sounds with specific visual elements, which I did not find suitable for my project. I preferred an approach that uses the whole structure of sound to create a complete image, as this approach is critical to my project's concept.

# System Design and Methodology

The pipeline begins with audio data, which is converted into a spectrogram, then processed using Neural Style Transfer (NST) to apply a painting style, resulting in a new image that blends the visual structure of the spectrogram with the style of the chosen painting.

## Pipeline

To collect the audio, users are given two options: they can either upload a pre-recorded audio file or use the built-in recording feature to capture sound directly. Once the audio is uploaded or recorded and the "generate" button is pressed, the *librosa* Python library, which is commonly used for audio analysis, is used to load the audio file. A Short-Time Fourier Transform (STFT) is then applied, which breaks the audio into small segments to analyze its frequency content over time. The amplitudes are then converted to decibels, to help compress the sound intensity into a more manageable range, enhancing the visibility of quiet sounds. The resulting spectrogram is visualized and saved as an image for the user to download.

Once the spectrogram is generated and displayed for the user, they can select one painting from a selection of five to use as the style for the new image, which will then be generated using Neural Style Transfer. The five painting options provided are Vincent van Gogh's *The Starry Night*, Edvard Munch's *The Scream*, Hokusai's *The Great Wave Off Kanagawa*, Claude Monet's *Impression, Sunrise*, and Salvador Dalí's *The Persistence of Memory*.

Once a painting is selected, it is transformed into a piece of art using the created spectrogram as the content image and the painting as the style image. Using a pre-trained arbitrary image stylization model, it blends the two to produce an original artwork that reflects both the structure of the sound and the aesthetic of the painting.

Before Neural Style Transfer is performed, both the content image (spectrogram) and the style image (painting) are resized to a smaller resolution of 256 by 256 pixels to comply with the model's input requirements. Each image also undergoes creative alterations: random Gaussian noise with a strength between 0.05 and 0.08 is added to simulate texture, pixel shifting ranging from -20 to 20 pixel is applied, and a ghosting effect with a transparency between 0.3 and 0.6 is also applied. These effects are applied to both the content and style images to introduce visual unpredictability, ensuring that the final composition is unique every time.

Once all preprocessing and the style transfer are completed, the user will be able to view and save the newly created artwork.

### User Interface

The user interface for this project is built using Gradio, a Python library that enables the creation of interactive interfaces for machine learning models. Users can either upload a pre-recorded audio file or use the built-in recording feature to capture sound, with both options separated by tabs at the top of the page. Once the spectrogram is generated, it is displayed to the right of the audio options. Below, users can select one of five paintings, with each option displayed in a row of images beneath the selection buttons. The final artwork is shown in its own row at the bottom of the page and can be viewed and saved directly through the interface, ensuring a seamless and user-friendly experience.

## Results and Analysis

Figure 1 presents an example of the Neural Style Transfer process, where audio of street sounds is first converted into a spectrogram and then transformed into a unique visual composition inspired by Vincent van Gogh's The Starry Night through style transfer. Since the output is dependent on the input audio, each transformation will result in a different visual representation.

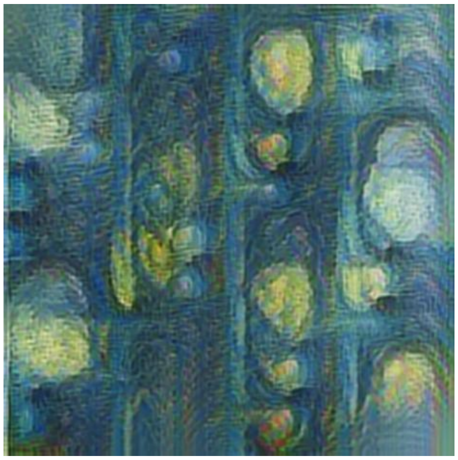

Figure 1. Neural Transfer Style result generated using Vincent Van Gogh's *The Starry Night* as the style image.

### Results

The results of the Neural Style Transfer process reveal how audio can be transformed into a piece of compelling art. When generating an image using Neural Style Transfer, the model successfully applied the visual style of the selected painting to the spectrogram image, preserving dominant colors and recognizable artistic features such as Van Gogh's swirling strokes, as seen in the yellow and blue swirls in Figure 1. This repetition of features creates a unique blend of audio structure and artistic style.

Each painting style produced visually distinct outputs. For example, the soft colours and flowing lines of *Impression, Sunrise* produced a dreamlike composition as seen in Figure 2. In contrast, *The Great Wave off Kanagawa* produced bolder contrasts and a sense of movement, created through the transfer of the wave feature from the painting onto the spectrogram image, as seen in Figure 3

Creative alterations like Gaussian noise, pixel shifting, and ghosting further contributed to the unique aesthetic in each output. These alterations ensured variation, reinforcing the generative nature of the process.

Some visual irregularities are also present in the output images. Vertical lines are noticeable across most of the results, especially in lighter areas as seen in Figure 3. In some cases, color distortion also appears along the edges of the results, especially near the bottom, adding unintended textures and shifts in colours.

Despite these imperfections, the results demonstrate a successful fusion of sound and art, producing unique outputs that vary depending on the audio input and selected painting.

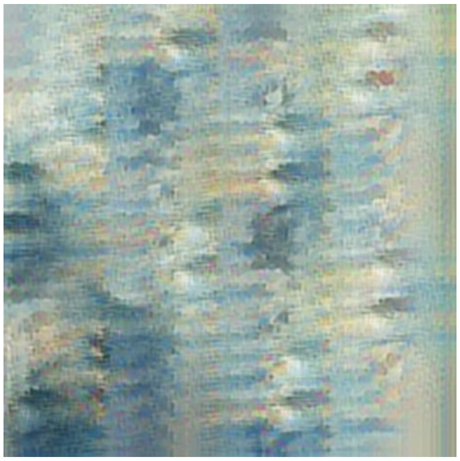

Figure 2. Neural Transfer Style result generated using Claude Monet's *Impression, Sunrise* as the style image.

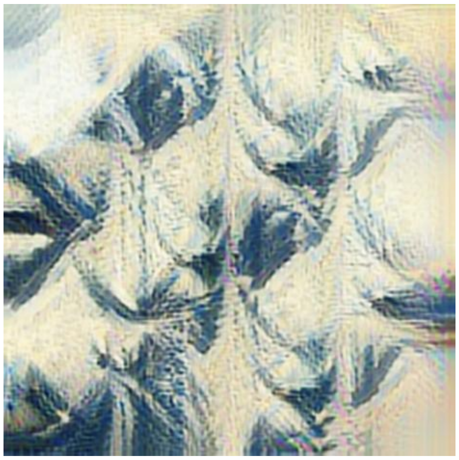

Figure 3. Neural Transfer Style result generated using Hokusai's *The Great Wave off Kanagawa* as the style image.

### Analysis

The presence of vertical lines in the neural style transfer results were likely caused by the resizing process, particularly because the aspect ratio of the image wasn't maintained when downscaling. When an image is resized without preserving its original aspect ratio, it can become distorted, which in this case resulted in the unwanted vertical lines. This occurs due to the independent scaling of the width and height, causing distortion specifically in where the high-frequency sounds are visualized. Since these frequencies are shown vertically, reducing the resolution emphasizes their form, resulting in more pronounced lines. Additionally, the color distortions near the edges, especially at the bottom of the images, result from how the Neural Style Transfer model interprets regions of high intensity or contrasting colours in the original spectrogram. These areas, when combined with the transferred features, can lead to exaggerated color blending or bleeding, especially in areas that lack visual features.

The distinctive features in the results of the style transfer highlight the model's sensitivity to both the visual characteristics of the chosen painting and the structure of the input spectrogram. The consistent transfer of distinct features such as Van Gogh's swirling brushstrokes suggests that the model prioritizes texture and color over strict spatial accuracy. This reinforces my understanding that NST captures style as a set of visual patterns rather than the exact brushstrokes and lines.

The creative additions like Gaussian noise and pixel shifts not only introduced variations to the output but may have also influenced how the style was applied. These changes likely disrupted the model's perception of the structure of both the image and the painting, creating more abstract and unpredictable outcomes. This unpredictability aligns with my creative goals of the project, turning audio spectrograms into unique visuals.

### Creative and Technical Implications

This project highlights the potential of Neural Style Transfer as both a creative tool and a method for exploring sound and visual art. The diversity in output, produced by applying different painting styles demonstrates how powerful style transfer can be as a generative medium. Each artwork has its own distinct style in terms of colours, textures, and brushstrokes, and Neural Style Transfer is able to effectively capture and transfer these unique visual elements onto a new image. This fusion of content and style opens new possibilities for creative expression, allowing for audio to take on the characteristics of iconic paintings, merging sound and visual art.

Technically, this process also reveals some challenges and considerations. Image preprocessing, such as resizing, can create unintentional distortions that affect both the structure and aesthetics of the final output. For example, improper aspect ratio handling can create unwanted visual artifacts, such as vertical lines or stretched features. These issues emphasize the importance of carefully preparing input data, especially when working with non-traditional inputs like spectrograms.

## Conclusion

This project demonstrates the transformative potential of Neural Style Transfer as a tool for combining sound and visual art. By successfully applying the visual styles of iconic paintings to the structures of spectrograms, this project not only showcases the potential of style transfer algorithms but also explores a new creative possibility in multi-sensory art generation.

Despite some technical challenges, such as colour distortions and visual irregularities, the results highlight the power of neural networks in generating artistic interpretations of complex data. These imperfections, rather than hindering the outcome, contribute to the unpredictability of the process, reinforcing its artistic value.

## Future Possibilities

Echoes of Art was created with the intention of being an interactive artwork. However, before it can be publicly displayed, the issues of colour distortions, noticeable vertical lines, and resolution limitations need to be addressed. The current model being used is limited to a 256 by 256 pixel output, and either training a model that supports higher resolution or incorporating automatic upscaling would significantly enhance this project's visual quality. One important consideration when using a new model is speed. The pretrained model was primarily selected for its speed and accuracy in style transfer, ensuring a balance between performance and output quality.

## Acknowledgements

I would like to extend my gratitude to everyone who supported me throughout the development of this project. Your encouragement and guidance have been invaluable.

I am also grateful to the developers and communities behind the Python libraries: TensorFlow, Gradio, and Librosa. Without these tools and resources, this project would not have been possible as they provided the technical foundation for my work. Special thanks to Google Colab for offering an accessible and powerful platform to develop and host the application.

A key component of this project was made possible by the Neural Style Transfer model developed by Google and made available through TensorFlow. Their work played a pivotal role in bringing this project to life.

## Ethics Statement

This project was developed with a strong consideration for ethical practices with the use of Artificial Intelligence and copyright art. All audio data used during testing was either recorded by the user or obtained from copyright-free public sources. The five famous paintings used are in the public domain and do not infringe on any intellectual property laws. This system does not collect or store any personal data from users, preserving their privacy. Additionally, the intent of this project is artistic and exploratory rather than commercial, aiming to expand the creative possibilities of sound-to-image translation without exploiting or misrepresenting the work of original artists or users.

## References

[1] Ziyang Chen, Daniel Geng, and Andrew Owens. "Images that Sound: Composing Images and Sounds on a Single Canvas." *Advances in Neural Information Processing Systems* 37, accessed April 12, 2025, https://doi.org/10.48550/arXiv.2405.12221

[2] Kim Sung-Bin, Arda Senocak, Hyunwoo Ha, Andrew Owens, and Tae-Hyun Oh. "Sound to Visual Scene Generation by Audio-to-Visual Latent Alignment." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, accessed April 12, 2025, https://doi.org/10.48550/arXiv.2303.17490

[3] Yonggai Zhuang, Yuhao Kang, Teng Fei, Meng Bian, and Yunyan Du. “From Hearing to Seeing: Linking Auditory and Visual Place Perceptions with Soundscape-to-Image Generative Artificial Intelligence.” *Computers, Environment and Urban Systems* 110, accessed April 12, 2025, https://doi.org/10.1016/j.compenvurbsys.2024.102122

## Bibliography

Chen, Ziyang, Daniel Geng, and Andrew Owens. "Images That Sound: Composing Images and Sounds on a Single Canvas." *Advances in Neural Information Processing Systems* 37 (2025): 85045-85073.
https://doi.org/10.48550/arXiv.2405.12221

Sung-Bin, Kim, Arda Senocak, Hyunwoo Ha, Andrew Owens, and Tae-Hyun Oh. "Sound to Visual Scene Generation by Audio-to-Visual Latent Alignment." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6430-6440. 2023.
https://doi.org/10.48550/arXiv.2303.17490

Zhuang, Yonggai, Yuhao Kang, Teng Fei, Meng Bian, and Yunyan Du. “From Hearing to Seeing: Linking Auditory and Visual Place Perceptions with Soundscape-to-Image Generative Artificial Intelligence.” *Computers, Environment and Urban Systems* 110 (June 2024).
https://doi.org/10.1016/j.compenvurbsys.2024.102122