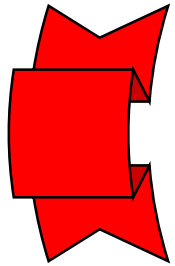


TräumerAI: Dreaming Music with StyleGAN

Dasaem Jeong, Seungheon Doh, Taegyun Kwon

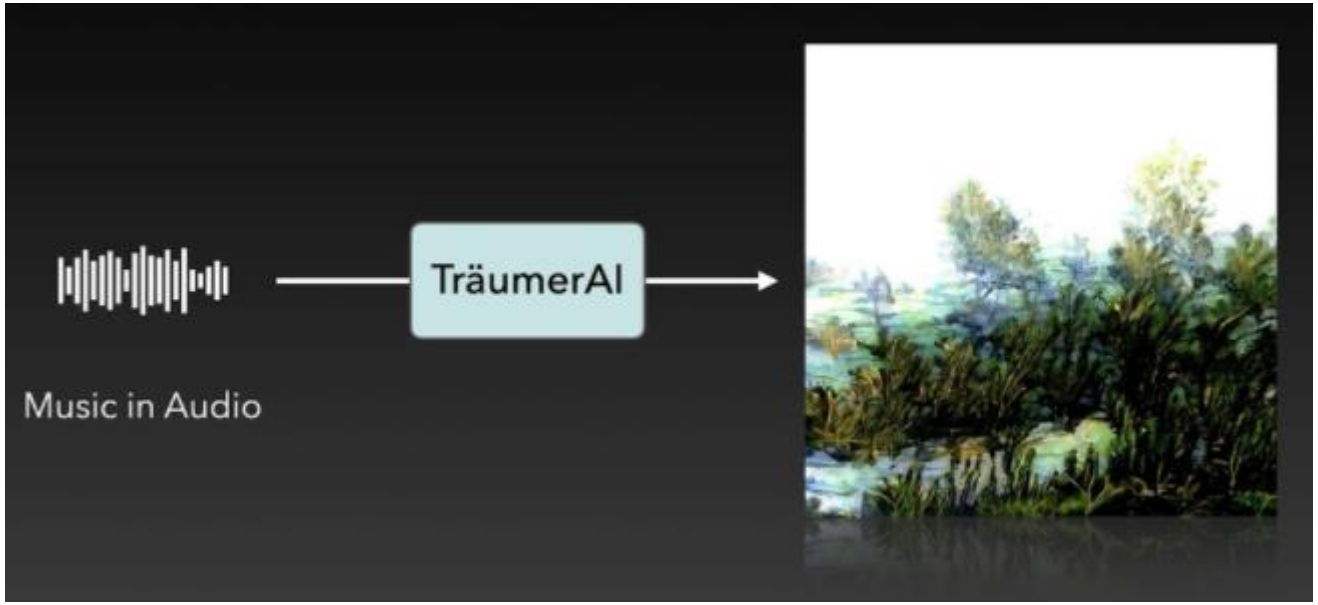
Supervisor：Prof. 陳自強 Oscal T.-C. Chen
Student：蔡承宏(Tsai Cheng Hong)
Student ID：609415074

Outline

- Introduction
- Overview
- System Implementation
- Conclusion
- Reference
- Demo video

Introduction

The goal of this paper **to generate a visually appealing video that responds to music with a neural network** so that each frame of the video reflects the musical characteristics of the corresponding audio clip. To achieve the goal, we propose a neural music visualizer directly mapping deep music embeddings to style embeddings of StyleGAN, named TräumerAI.

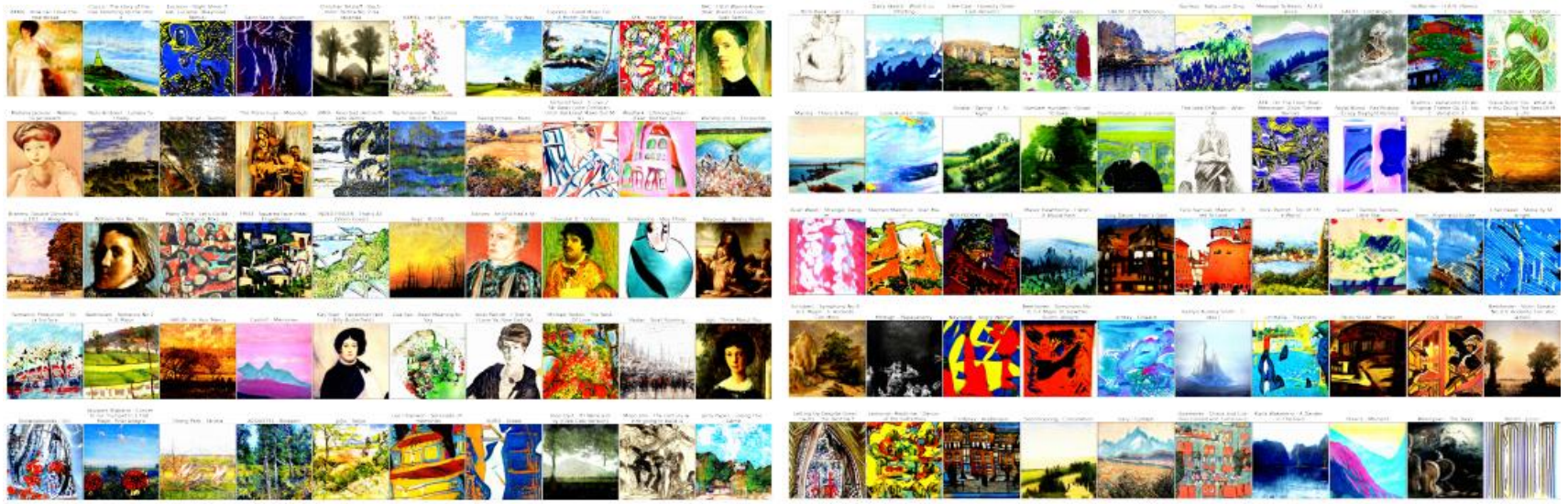


Big picture

Introduction

TräumerAI :

- A music auto-tagging model using CNN and StyleGAN2 pre-trained on WikiArt dataset.
- They manually labeled the pairs in a subjective manner.
- Selected an image that suits the music among the 200 StyleGAN-generated examples.
- trained a simple transfer function that converts an audio embedding to a style embedding.



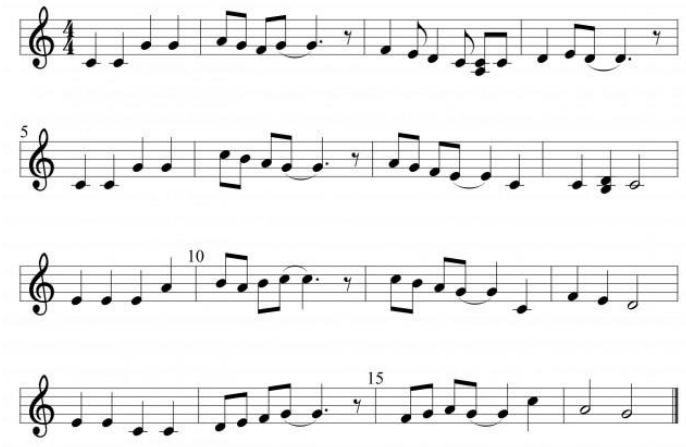
Selected images

Introduction

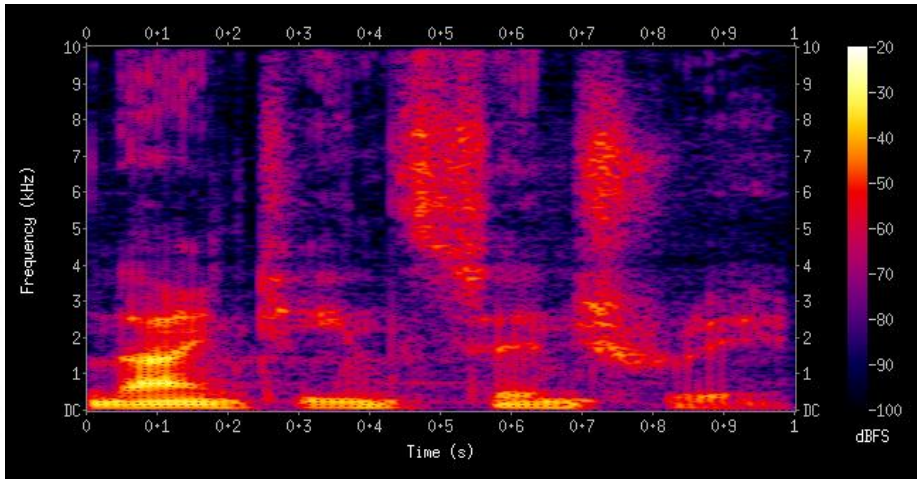
There are many commonly used visual representations for music.

Such as : Music notation, Spectrogram, Piano roll.

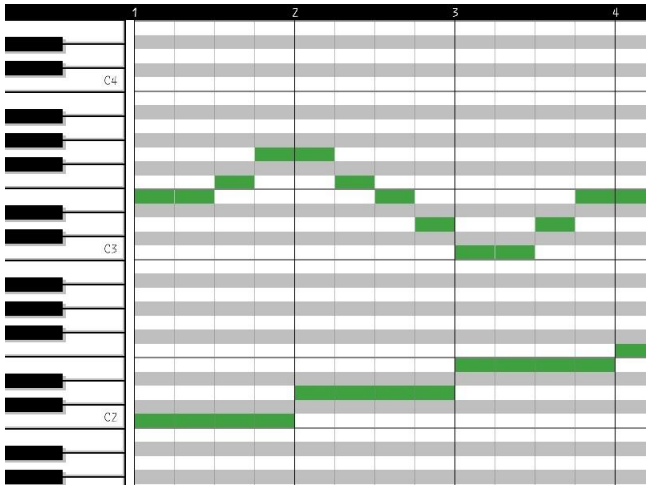
Music visualization can provide additional information via visual.



music notation



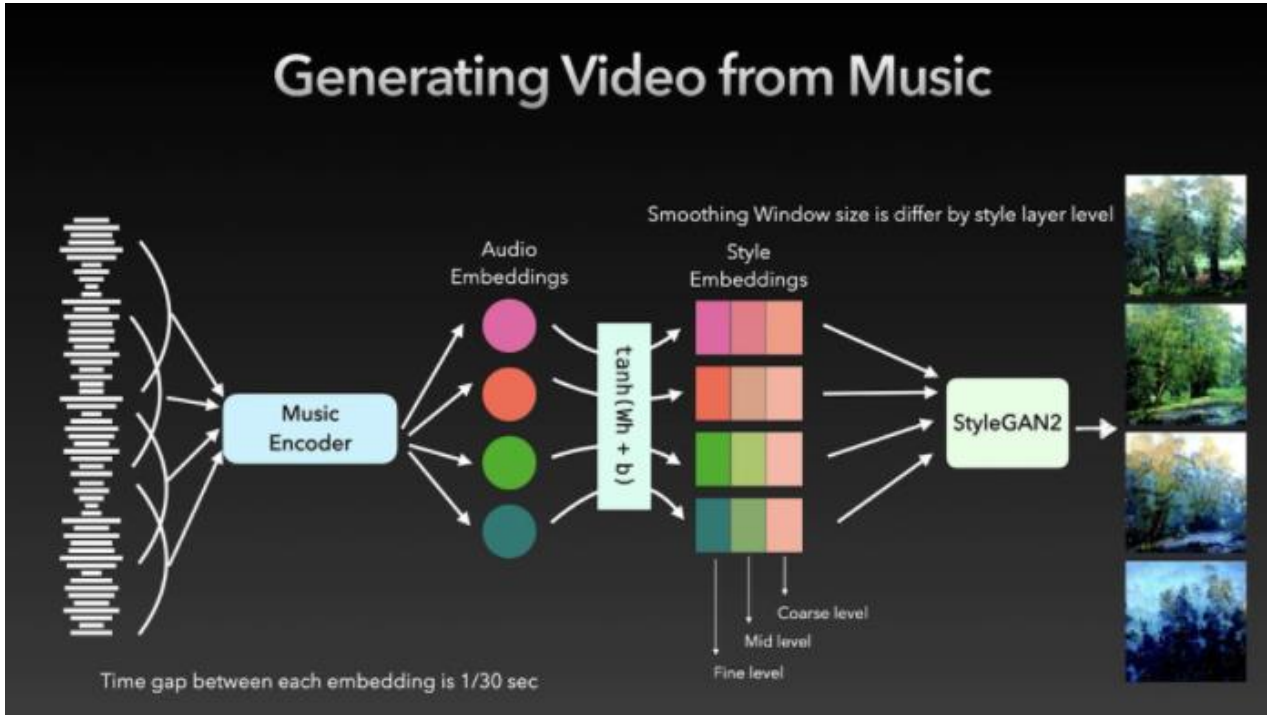
spectrogram



piano roll

Overview

Step 1. Music sementic embedding extraction
Step 2. Mapping music latent to visual latent
Step 3. Multi level smoothing

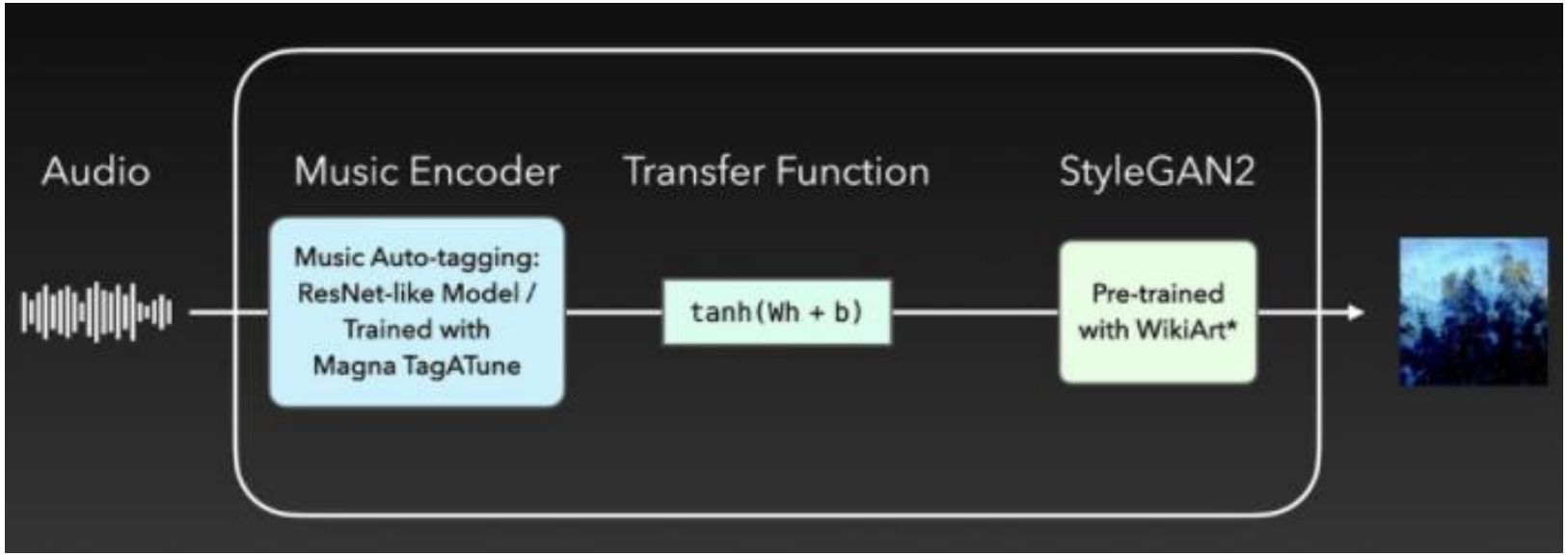


Structure of TräumerAI system

System Implementation

Audio Embedding

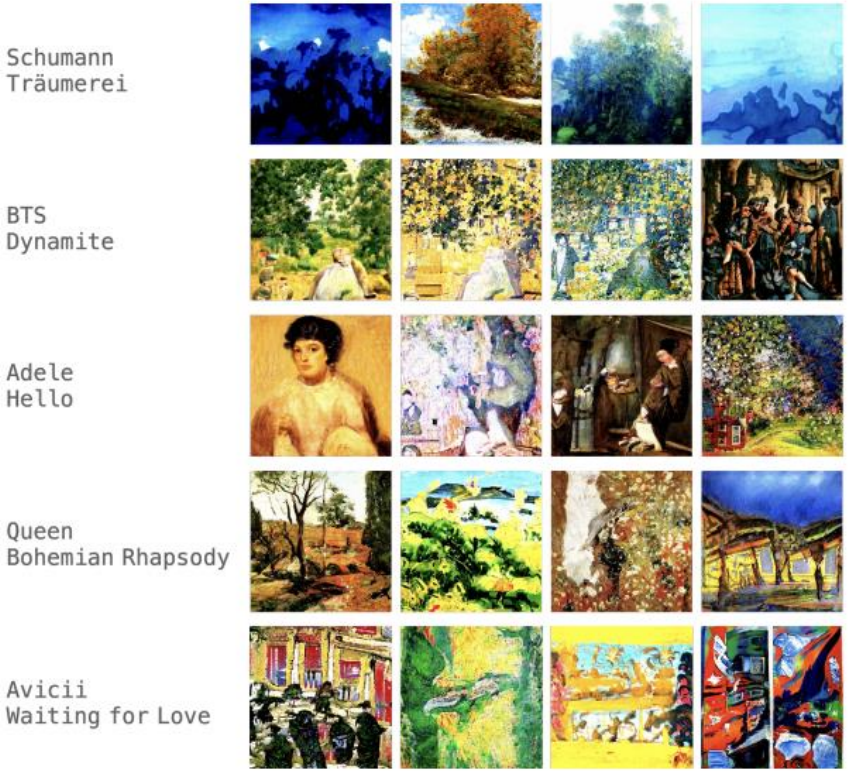
- Using a music auto-tagging model as a fixed music encoder, It is a deep fully convolutional network (FCN).
- Automatic music tagging is a multi-label binary classification task that aims to predict relevant tags for a given song.
- Used the output of the last CNN layer as an embedding of the audio.



System Implementation

Image Generator

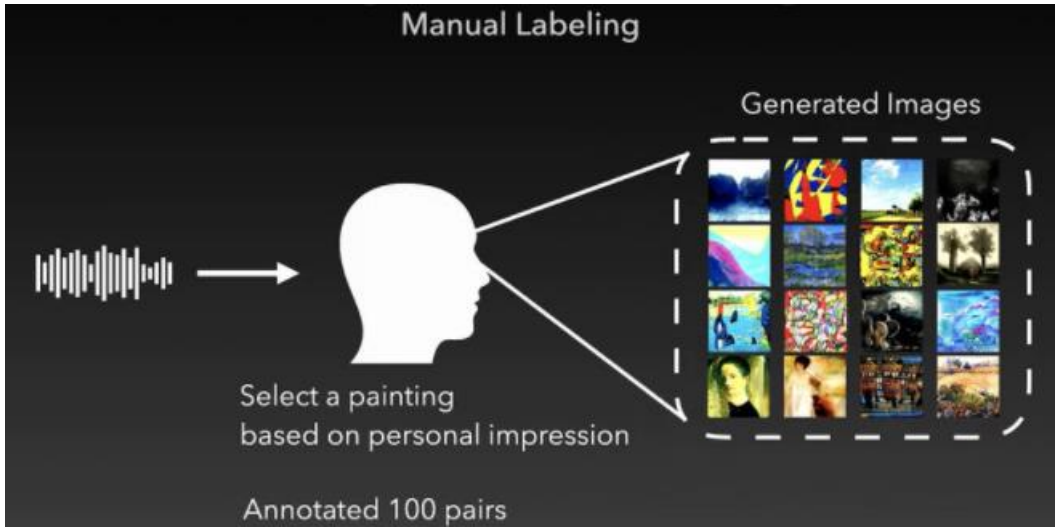
- Used StyleGAN to generate image.
- StyleGAN can provide high-resolution images of quality.

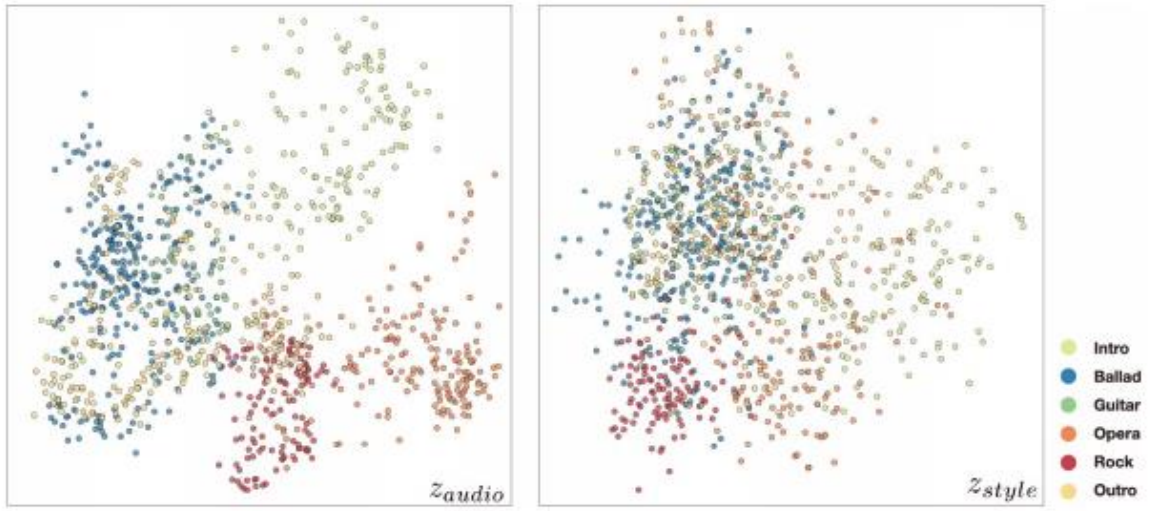


System Implementation

Manual labeling between Music and Image

- **Manual labeled the pairs in a subjective manner**, because we have no objective metric between musical and visual semantics.
- **Selected an image that suits the music** among the 200 StyleGAN-generated examples.
- If could not find an appropriate image, he generated another 200 images in random. The data covers various genres including classical, jazz, pop, ballad, R&B, new age, K-pop, J-pop, rock, electronic, hip hop, and trot.





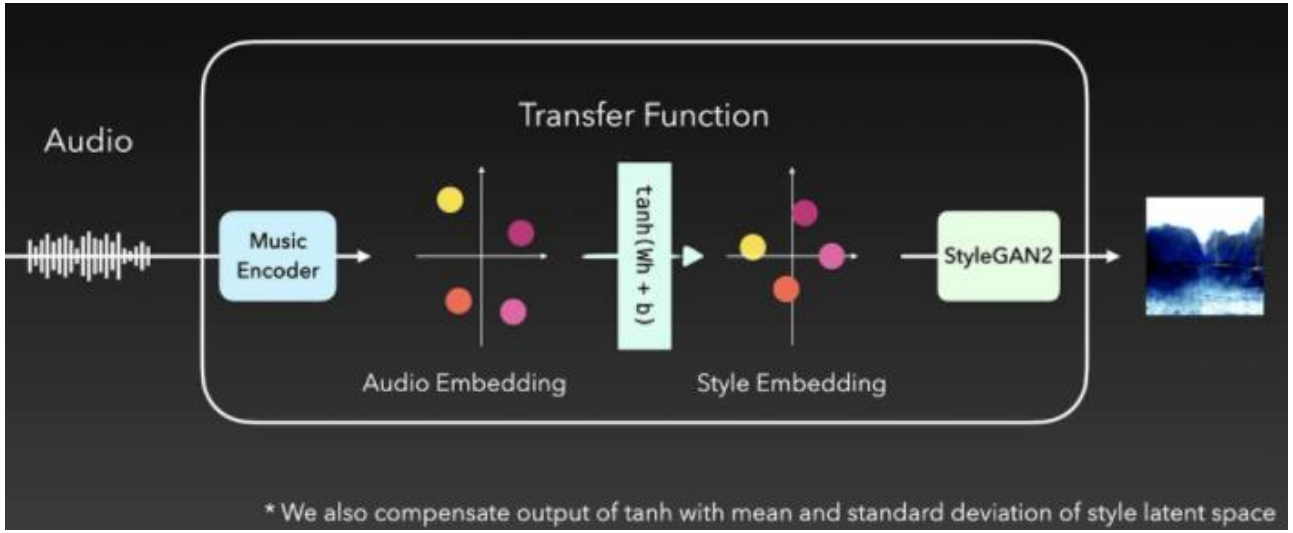
audio encoder extracts different audio embedding for each sub-genre of the music

System Implementation

Transfer function

- we trained a simple transfer function that converts an audio embedding z_{mu} to a style embedding w_{st}.
- This transfer function is similar to the one used for zero-shot learning (Semeantic Spcae) between words and images.
- With mean and deviation μ_{st}, σ_{st} of w_{st} from random sampled z of StyleGAN.

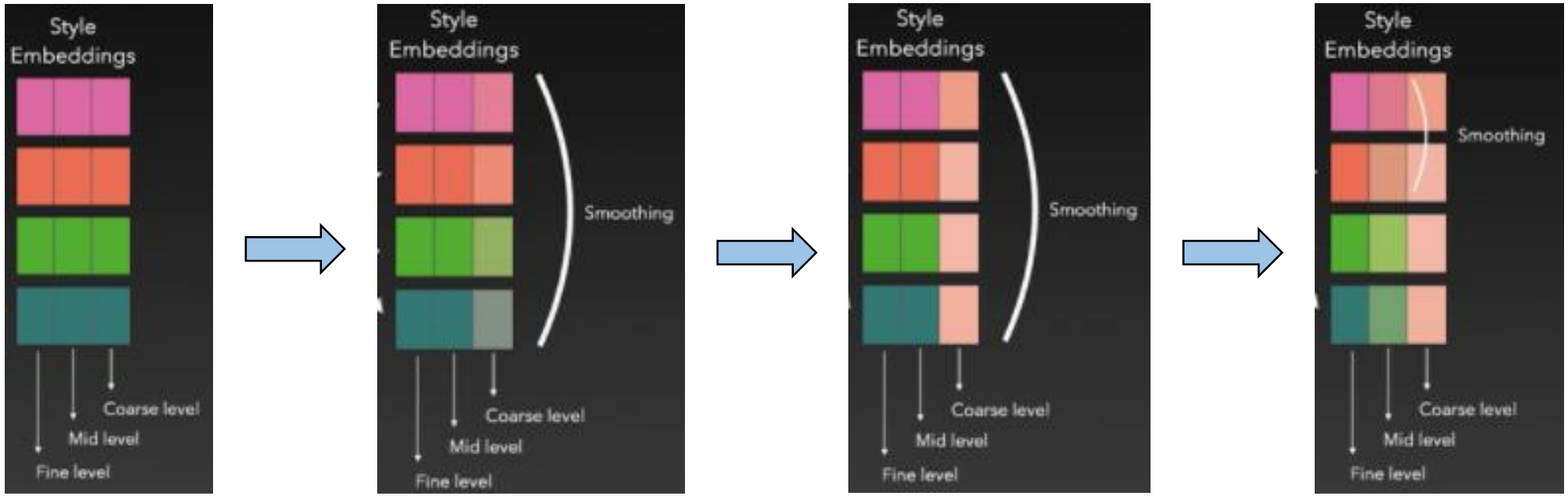
$$\mathcal{L}(w_{st}, z_{mu}) = \sum |w_{st} - (2\sigma_{st} \tanh(W z_{mu} + b) + \mu_{st})|$$



System Implementation

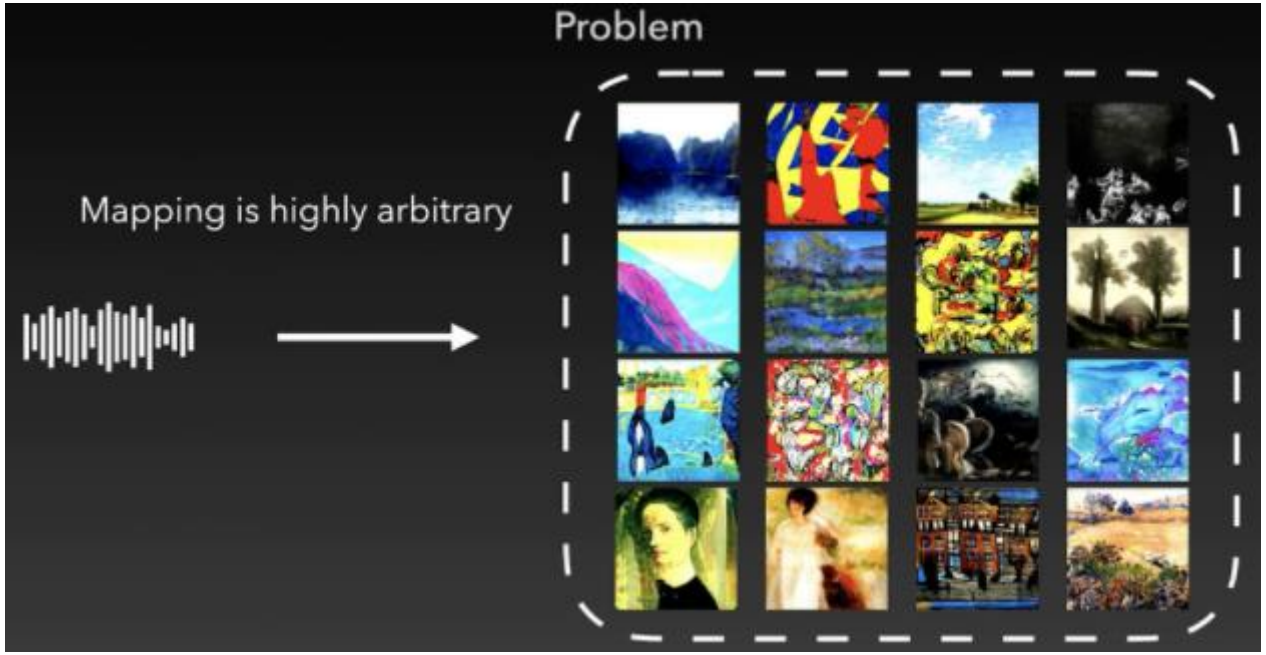
Video Generation

- Sampled the 30 audio embeddings per second so that each frame of video is generated from the corresponding audio embedding.
- Smoothing with an averaging window is applied to the style sequence to prevent the generated images from changing too rapidly.
- The window size differs by style hierarchy, so that coarse, middle, and fine styles are smoothed with a window of 3 sec, 2 sec, and 0.3 sec, respectively.



Conclusion

- Since the mapping between music and image is done in subjective pairs, the generated results are heavily biased by the annotator's preference on music.
- Implementing a personal version of the neural music visualizer is valuable.
- If used an active learning method that will significantly reduce the time for the labeling process.



Reference

1. Jeong, D., Doh, S., & Kwon, T. (2021). TräumerAI: Dreaming Music with StyleGAN. ArXiv, abs/2102.04680.
2. Won, M., Ferraro, A., Bogdanov, D., & Serra, X. (2020). Evaluation of CNN-based Automatic Music Tagging Models. ArXiv, abs/2006.00751.
3. Socher, R., Ganjoo, M., Manning, C.D., & Ng, A. (2013). Zero-Shot Learning Through Cross-Modal Transfer. NIPS.

Demo video

Song: Avicii – Waiting For Love

END