TräumerAI: Dreaming Music with StyleGAN

Dasaem Jeong
T-Brain X
SK Telecom
Seoul, South Korea
dasaem.jeong@sktbrain.com

Seungheon Doh **Taegyun Kwon**
Graduate School of Culture Technology
KAIST
Dajeon, South Korea
{seungheondoh, ilcobo2}@kaist.ac.kr

1 Introduction

Although music is usually regarded as an audio domain, there are many commonly used visual representations for music, such as music notation, spectrogram, and piano roll. Because the music visualization can provide additional information via visual, there have been various music visualization schemes with different purposes, such as to visualize the emotion of music by selecting photos[1], to implement an active listening interface by visualizing structure[2] or progress of music[3], or to create media art performance[4].

After the recent advances of generative models, there have been several works exploring the cross between music and visual domain using deep neural networks. An audio-reactive StyleGAN[5] was introduced, which navigate the latent space of StyleGAN with controlled speed based on audio features such as digital filtering outputs and Nsynth[6]. The limitation is that the starting images are manually selected for each generation, rather than automatically generated, and only the movement between images are controlled by the acoustic features. Also, selected audio features are focused on the loudness and timbral aspects of audio, the video does not match with high-level semantic features of music such as genre and mood.

Another recent work proposed a crossing between music and visual style based on artistic periods, such as mapping Debussy's music to French Impressionists' style[7]. The mapping based on the era provides objective shared labels and helps to avoid arbitrariness in pairing music and art. However, as the authors themselves pointed out, the era label is not sufficient enough to bridging between music and paintings. Also, the model generates a visual style rather than an image, so it demands an additional reference image to be style transferred.

Our goal is 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 [1].

2 System Implementation

2.1 Audio Embedding and Image Generator

We utilized a music auto-tagging model as a fixed music encoder, which is a short-chunk CNN with residual connection based on [8], and trained the model with MagnaTagATune dataset [9] and its top 50 tags. We used the output of the last CNN layer as an embedding of the audio.

For image generation, we used StyleGAN [10] due to its capacity of generating high-resolution images of quality. Our system was made from a public PyTorch implementation[2] of StyleGAN2 [11] and a pre-trained model that was trained with WikiArt Dataset[3].

[1]The code is available on https://github.com/jdasam/traeumerAI. We encourage the readers to watch our demo videos on https://jdasam.github.io/traeumerAI_demo/

[2]https://github.com/rosinality/stylegan2-pytorch

[3]https://github.com/pbaylies/stylegan2

4th Workshop on Machine Learning for Creativity and Design at NeurIPS 2020, Vancouver, Canada.

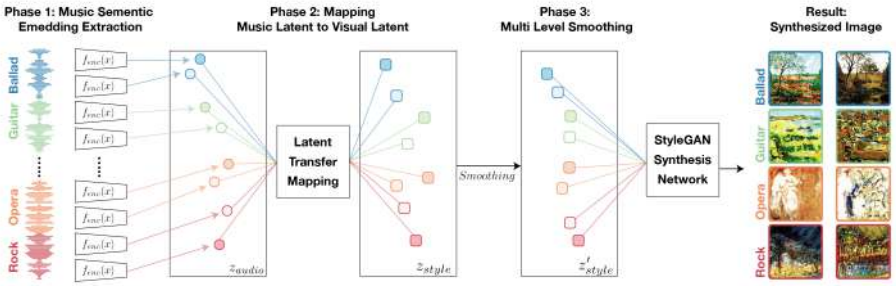

Figure 1: Structure of the proposed TräumerAI system. Images on the right are generated output of our system from Queen's "Bohemian Rhapsody"

2.2 Manual labeling between Music and Image

Rather than establishing an objective metric between musical and visual semantics, we manually labeled the pairs in a subjective manner. One of the authors listened to 100 music clips of 10 seconds long and selected an image that suits the music among the 200 StyleGAN-generated examples. If the annotator could not find an appropriate image, he generated another 200 images in random or based on a single image from the previous step. The data covers various genres including classical, jazz, pop, ballad, R&B, new age, K-pop, J-pop, rock, electronic, hip hop, and trot.

During the process, the annotator considered his emotional impressions such as arousal and valence, genre and era, and timbral characteristic such as instrumentation. We intentionally avoided mapping vocal sounds to a portrait unless the vocal is strongly dominant or any other instruments does not appear, because portraits have relatively low variance on visual style or perceptive emotion.

Based on the collected data, 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 between words and images in [12], with mean and deviation μ_{st}, σ_{st} of w_{st} from random sampled z of StyleGAN.

$$\mathcal{L}(w_{st}, z_{mu}) = \sum |w_{st} - (2\sigma_{st} \tanh(W z_{mu} + b) + \mu_{st})| \tag{1}$$

2.3 Video Generation

The system takes audio input and extracts a sequence of audio embedding. The sequence is converted to a sequence of styles, which is then generated as a sequence of images by StyleGAN. During the experiment, linear interpolation between coarsely sampled sequence(3 sec) results in abrupt acceleration. Therefore, we sampled the 30 audio embeddings per second so that each frame of video is generated from the corresponding audio embedding. Since our mapping conserves the temporal progress of embeddings, the progress of the video followed structural change of music. Also, 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.

3 Discussion

The generated video on Queen's "Bohemian Rhapsody", which can be segmented into six different sub-genres, shows that the mapping between audio and video makes a certain level of intra-segment similarity and inter-segment dissimilarity as presented in Figure 1.

Although exploring objective mapping between different domains is interesting, subjective mapping can still be a reasonable solution due to the subjective nature of art. Therefore, making a user interface for efficient data labeling between music and painting can be valuable for implementing a personal version of the neural music visualizer. Designing an easily navigable system for the StyleGAN latent space, and applying an active learning method that helps annotators efficiently cover various music or painting styles will significantly reduce the time for the labeling process, which are remained for the future work along with a quantitative evaluation.

Ethical Implications

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. Therefore, the system can generate bizarre or grotesque images from the music, which may deliver a biased impression on the music to viewers, thus may distort the original artist's intention or creativity.

References

[1] C.-H. Chen, M.-F. Weng, S.-K. Jeng, and Y.-Y. Chuang, "Emotion-based music visualization using photos," in *Proc. of International Conference on Multimedia Modeling*, Springer, 2008, pp. 358–368.

[2] M. Goto, "Active music listening interfaces based on signal processing," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, vol. 4, 2007, pp. IV–1441.

[3] D. Jeong and J. Nam, "Visualizing music in its entirety using acoustic features: Music flowgram," in *Proc. of the 2nd International Conference on Technologies for Music Notation and Representation (TENOR)*, 2016, pp. 25–32.

[4] R. Taylor, P. Boulanger, and D. Torres, "Real-time music visualization using responsive imagery," in *Proc. of 8th International Conference on Virtual Reality*, 2006, pp. 26–30.

[5] H.-H. Lee, D.-G. Wu, and H.-T. Chen, "Stylizing audio reactive visuals," in *NeurIPS 2019 Workshop: Machine Learning for Creativity and Design*, 2019.

[6] 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 *Proc. of the 34th International Conference on Machine Learning (ICML)*, PMLR, 2017, pp. 1068–1077.

[7] C.-C. Lee, W.-Y. Lin, Y.-T. Shih, P.-Y. P. Kuo, and L. Su, "Crossing you in style: Cross-modal style transfer from music to visual arts," *arXiv preprint arXiv:2009.08083*, 2020.

[8] M. Won, A. Ferraro, D. Bogdanov, and X. Serra, "Evaluation of CNN-based automatic music tagging models," in *Proc. of 17th Sound and Music Computing (SMC)*, 2020.

[9] E. Law, K. West, M. I. Mandel, M. Bay, and J. S. Downie, "Evaluation of algorithms using games: The case of music tagging.," in *ISMIR*, 2009, pp. 387–392.

[10] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proc. of the IEEE conference on Computer Vision and Pattern Recognition*, 2019, pp. 4401–4410.

[11] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of stylegan," in *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 8110–8119.

[12] R. Socher, M. Ganjoo, C. D. Manning, and A. Ng, "Zero-shot learning through cross-modal transfer," in *Advances in neural information processing systems*, 2013, pp. 935–943.

Supplementary

Example of the Labeling

The selected music and corresponding images during the annotation process are presented in Figure 2. The images were selected from random generation of StyleGAN without truncation to attempt to fully exploit its diversity in expression. Some images like No. 28 and No. 67 are extremely distant from other images in terms of style latent vector, as we did not take into account how extreme the image is in the style latent space. The annotation process took about 10 hours, and we did not modified any annotation after or during the process. To compensate these outliers, we used L1 loss and tanh non-linearity. To be clear, the 100 labeled music clips do not include any music of the artists who are selected for video demonstration.

The annotation process became time consuming when to cover extremely different genre like Korean trot. If an annotator limits the genre of music and style of image in certain level annotation process can become much faster. Therefore, We expect that other users can make their own version of audio-visual mapping by their preference in shorter time.

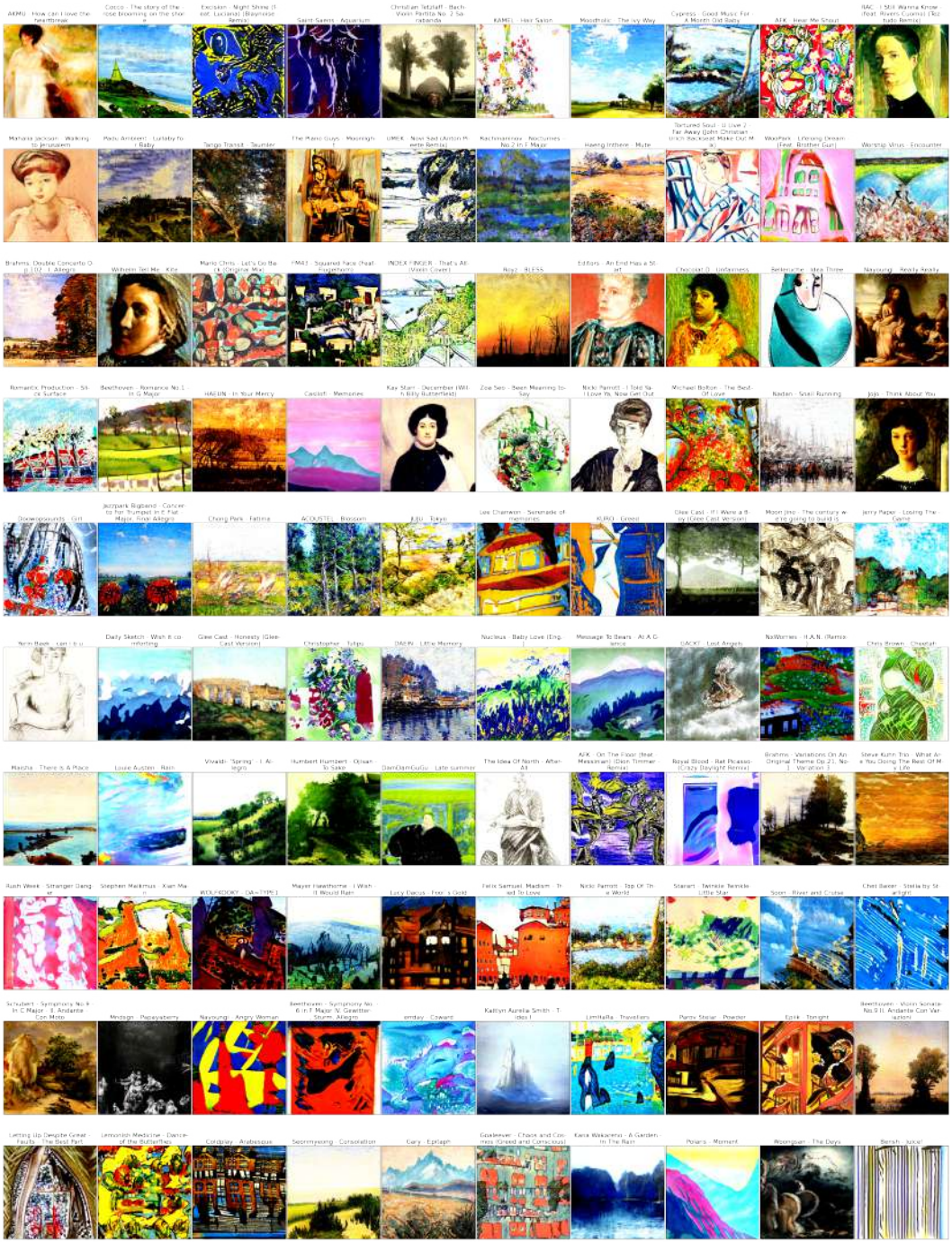

Figure 2: List of 100 music and corresponding images that were paired during the annotation. Some of the Korean names and titles are translated by the author.

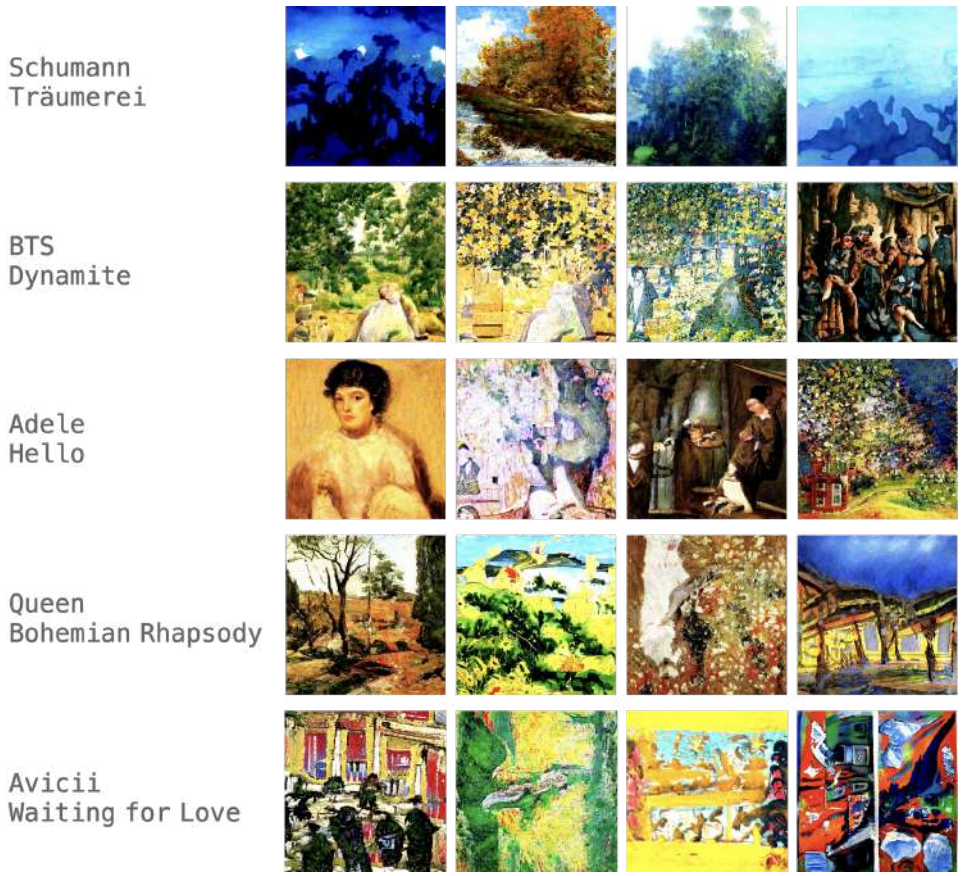

Figure 3: Snapshots from example videos. Images in the same row are generated from different segments of the same music

Example of Generated Images

Figure 3 shows examples of generated images. Again, we encourage the readers to watch videos on https://jdasam.github.io/traeumerAI_demo/, since how the image is changed throughout the music is the main part of our contribution.

Example of Audio and Style Embedding Trajectory

Figure 4 demonstrates how audio embeddings and its corresponding mapped style embeddings changes as music progresses, represented in 2D PCA. The presented result shows that our audio encoder extracts different audio embedding for each sub-genre of the music, and also that the mapped style vectors are conserving the tendency in certain level.

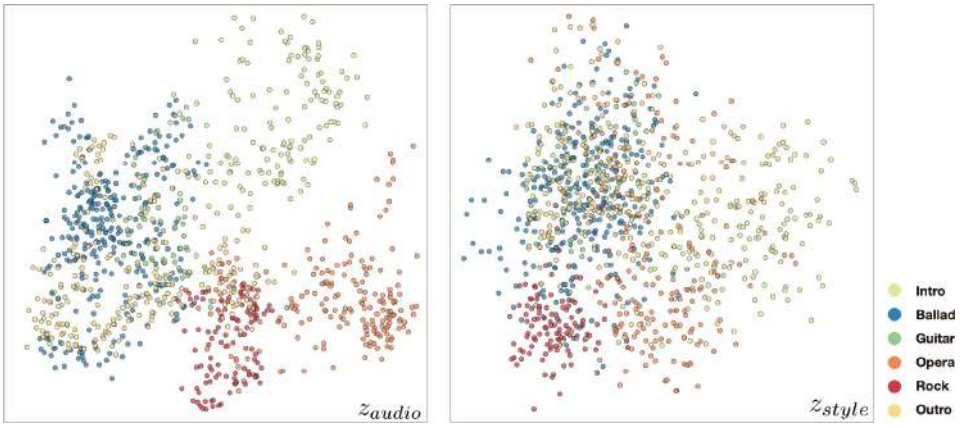

Figure 4: 2D PCA of audio embeddings and mapped style embeddings from Queen's "Bohemian Rhapsody". Each point represents 3.7 seconds of audio clip, which is sampled for every one third second