VIS2MUS: EXPLORING MULTIMODAL REPRESENTATION MAPPING FOR CONTROLLABLE MUSIC GENERATION

Runbang Zhang¹, Yixiao Zhang², Kai Shao¹, Ying Shan³, Gus Xia^{1,4}

Music X lab, NYU Shanghai
Centre for Digital Music, Queen Mary University of London
Tencent Inc. ⁴ MBZUAI

ABSTRACT

In this study, we explore the representation mapping from the domain of visual arts to the domain of music, with which we can use visual arts as an effective handle to control music generation. Unlike most studies in multimodal representation learning that are purely data-driven, we adopt an *analysis-by-synthesis* approach that combines deep music representation learning with user studies. Such an approach enables us to discover *interpretable* representation mapping without a huge amount of paired data. In particular, we discover that visual-to-music mapping has a nice property similar to *equivariant*. In other words, we can use various image transformations, say, changing brightness, changing contrast, style transfer, to control the corresponding transformations in the music domain. In addition, we released the Vis2Mus system as a controllable interface for symbolic music generation. ¹

Index Terms— Multimodal representation learning, Controllable music generation

1. INTRODUCTION

Multimodal representation learning aims to bridge the heterogeneity gap between different modalities [1]. Recently, multimodal representation learning has attracted widespread interest in applications such as cross-modal retrieval [2, 3], and cross-modal generation [4, 5, 6]. For audio, some deep learning-based work has progressed on tasks such as audio captioning [7], speech synthesis [8], and music description [9, 10, 11].

In this paper, we follow this research path and aim to explore the representation mapping from image to music. We believe that a better understanding of visual-to-music mapping will not only shed light on the theories of machine learning (especially transfer learning) and cognitive science (especially synesthesia) but also have great practical values on music information retrieval and controllable music generation. For most people, music concepts, such as chords and texture, are quite abstract and it is difficult to use these concepts to

directly control music generation. On the contrary, images are more "concrete" and can be used as an intuitive handle to preview and guide music generation once a multimodal mappings is established.

Due to insufficient available paired image-music data, existing data-driven approaches [12, 13] have difficulty in learning a mapping directly from data and generate high-quality music from images. To address this problem, we resort to *analysis-by-synthesis* methodology [14] and develop a method that combines representation learning with user subjective evaluations. In specific, we i) propose several transformations that can be applied to both music and image representations, ii) synthesize image-music pairs data using deep generative models by applying the proposed transformations, and iii) conduct user studies to evaluate the synthesized pairs to explore properties of visual-to-image mappings.

We find that visual-to-music mapping has a nice property analogous to equivariant. This property enables us to use various image transformations (including changing brightness, changing contrast and style transfer) to control the corresponding transformations in the music domain. In addition, we release the Vis2Mus system, which applies our approach to a large amount of paired image and music as a controllable interface for symbolic music generation.

2. METHODOLOGY

Given an image, the problem of "what is the ideal corresponding music" may seem too difficult or intractable, and different people probably have different answers. In this paper, we aim to solve a *conditioned* simplification of the problem which is more accessible, in which we: 1) first assume an image-music pair, 2) then modify the image by changing some features, and 3) finally ask "how we should change the music accordingly to match the modified image."

Such problem setup is also graphically represented in Figure 1, in which we can regard the given (conditioned) imagemusic pair as an "anchor" to control or judge any modification or variation based on it. Formally, let v denote the visual input, m denote the music output, and function f be the con-

¹GitHub repo: https://github.com/ldzhangyx/vis2mus.

ceptual visual-to-music mapping. Our problem setting is that given a specific pair m=f(v) and a certain transformation g such that v'=g(v), what is the corresponding music m', i.e., f(v')?

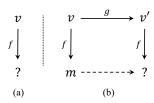


Fig. 1. An illustration of problem setup: (a) unconditioned multimodal mapping Vs. (b) conditional multimodal mapping. Our study focuses on the scenario of (b).

We develop an analysis-by-synthesis method that combines deep music generation with user studies. Inspired by Kandinsky's statements on synesthesia[15], our hypothesis is that f has an isomorphic property with respect to a certain transformations that can be applied in both visual and music domains. Formally, if m=f(v) and v'=g(v), then m'=g(m). In other words, given a pair (v,m), we can infer m' without knowing the exact form of f.

Under this hypothesis, we first propose several concrete forms of g. Then, we synthesize the corresponding m' via deep music generation. Finally, we conduct user studies to see whether users can distinguish m' from other candidates. In the rest of this section, we introduce the deep generative model and how to apply g on its latent representation in section 2.1, discuss several particular forms of g in section 2.2 and section 3.3, followed which we present the user study and experimental results in section 2.4 and section 2.5.

2.1. The Backend Deep-generative Model

In order to generate music from the latent space, we use a pretrained polyphonic music representation learning model (the bottom half of Figure 2) [16], which can learn disentangled chord and texture representations from music, as our backend model. Note that "texture"(as in English) can refer to both visual and musical features, and our hypothetical cross-modal transformation g acts on the latent texture feature.

The music model uses two separate encoders, a texture encoder and a chord encoder, to learn corresponding latent representations, $h_{\rm txt}$ and $h_{\rm chd}$. In specific, the texture encoder takes a quasi-piano-roll input and consists a CNN layer followed and a bi-directional GRU. The chord encoder takes a chord progression input and is implemented by a bi-directional GRU. The model also uses a PianoTree decoder [17] to reconstruct the music from the two representations.

The upper half of Figure 2 shows an example of image transformation. We let the same transformation g acts on both image v and the feature map $z_{m,\mathrm{txt}}$, the intermediate output of the CNN module of the texture encoder. The rationale

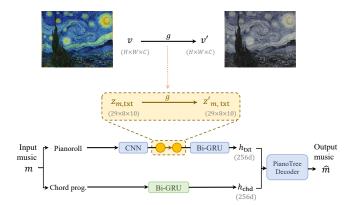


Fig. 2. An illustration of the polyphonic disentanglement model and how to apply cross-modal transformation g to the latent texture feature.

of such design is that the intermediate representation of the music texture $z_{\text{txt},m}$ is a two-dimensional feature map, which makes it easy to perform various image-like operations on it, such as changing the "brightness" and "contrast" of the feature map. Finally, The manipulated feature map $z'_{\text{txt},m}$ is sent back to the music model to reconstruct a new piece of music with the transferred representation.

2.2. Brightness and Contrast Transformation

2.2.1. Brightness Transformation

The brightness transformation g_b can be regarded as a pixelwise function. Let x be any channel (a matrix) of the z_m or v, then for every element (the pixel value) of x_i , its brightness transformation is defined as:

$$x_i' = g_b(x_i) = x_i + \alpha_b/2 \cdot \max(x) \tag{1}$$

where α_b is a hyperparameter. Figure 4 shows an example of synthesizing a new image-music pair (v', m') based on an existing pair (v, m) using brightness transformation.

2.2.2. Contrast Transformation

Similarly to brightness transformation, the contrast transformation g_c is defined as:

$$x_i' = g_c(x_i) = \overline{x} + (x_i - \overline{x}) \cdot (1 + \alpha_c) \tag{2}$$

where α_c is a hyperparameter. Figure 5 shows an example of synthesizing a new image-music pair (v', m') based on an existing pair (v, m) using contrast transformation.

2.3. Style Transfer

Besides changing brightness and contrast, we also consider style transfer as a special form of transformation. In general,

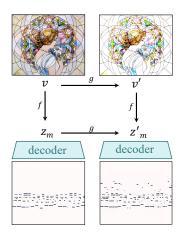


Fig. 3. An illustration of applying cross-modal brightness transformation g_b on image v and music m, respectively, in which (v, m) is a matched image-music pair.

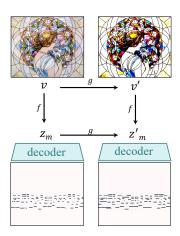


Fig. 4. An illustration of applying cross-modal contrast transformation g_c on image v and music m, respectively, in which (v, m) is a matched image-music pair.

style transfer is achieved by fusing a content element and a style element. In the image domain, people usually regard image contour as content and a colored texture as style [18]. Correspondingly, we regard melody contour as content and polyphonic texture as style in the music domain.

Formally, let $z=(z_c,z_{\rm txt})$ be the representation of an image or a piece of music, where z_c and $z_{\rm txt}$ represent the corresponding contour and texture, respectively. The cross-modal style transfer operation g_s generates the latent representation of a new image or music:

$$z' = g_s(z) = g_s(z_c, z_{\text{txt}})$$
 (3)

where g_s simply concatenate contour and texture representation.

Figure 5 shows an example of synthesizing a new imagemusic pair (v', m') based on (v_1, v_2) and (m_1, m_2) . Here, v_1

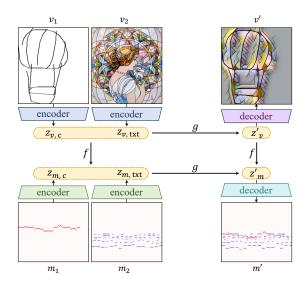


Fig. 5. An illustration of applying cross-modal style transformation g on the images (v_1, v_2) and the music pieces (m_1, m_2) , respectively, in which (v_1, m_1) and (v_2, m_2) are two matched image-music pairs.

is a sketched contour, v_2 is a styled image (with colored texture), m_1 is a melody contour, and m_2 is a styled accompaniment (with polyphonic texture). The encoder and decoder of the image style transfer part come from [19], while the encoder and decoder of the music style transfer part come from [20, 16]. In specific, the compositional style transfer algorithm first adjusts the pitch register and the tempo of the melody and accompaniment, respectively, to let them match each other. Then, our backend deep-generative model (as described in section 2.1) performs chord style transfer [16] to the accompaniment by substituting the latent representation of original chords with the latent representation of melody's chords. In this way, the music texture is fused into the melody, just as the style of the image is fused into the sketch.

2.4. User Study Design

The user study contains three parts: 1) brightness transformation user study, 2) contrast transformation user study, and 3) style transfer user study. The brightness transformation user study and the contrast transformation user study have the same three steps:

- 1. **Preview**: The user is asked to view and listen to an image-music pair;
- 2. Transformation: The image is transformed in terms of brightness or contrast, while the music is altered both by the same transformation and the inverse transformation, producing two new pieces of music. E.g. if the image is transformed to be brighter, the inverse transformation means to make the latent representation of

the paired music dimmer;

 Test: The user is asked to select which of the two new pieces better matches the transformed image. Therefore, the corresponding random guess baseline accuracy is 50%.

The style transfer user study also has 3 steps:

- Preview: The user is asked to view and listen to four image-music pairs: two sketch-melody pairs and two image-accompaniment pairs.
- 2. Transformation: A synthesized image is shown, which is transformed using a randomly selected one out of the two sketches as the content and a randomly selected one of the two images as the style. At the same time, four pieces of synthetic music are displayed, generated by all possible combinations of the two melodies and the two accompaniments shown in the first step using compositional style transfer.
- 3. **Test**: The user is asked to select one synthetic music from the four that perceptually best matches the synthesized image. Therefore, the corresponding random guess baseline accuracy is 25%.

2.5. Experimental Results

For the brightness user study and the contrast user study, the gender distribution of all subjects was 61% male and 39% female, and the music level distribution was 36% amateur, 39% intermediate, and 25% professional; a total of 24 and 22 responses were collected respectively. For the style transfer user study, the gender distribution was 76% male and 24% female, and the music level distribution was 39% amateur, 46% intermediate, and 15% professional; a total of 61 responses were collected. Table 1 shows the results of three user studies, with user selection accuracy significantly higher than baseline random guesses in all tests.

Acc (%)	Brightness	Contrast	Style
Random	50.0	50.0	25.0
Ours	62.5*	68.2*	60.6*

Table 1. Evaluation results of the transformation experiments. *: The accuracy outperforms random guesses with p < 0.05 on binomial test.

This result indicates that given an image-music pair (m,v) and a transformed image v' with respect to brightness, contrast or style, humans can tell the corresponding music m' that is "brighter"/"dimmer", with more/less "contrast", or style transformed without any training. The commutative cross-modal transformations do exist, though not for everyone, yet statistically significant. Such property also enables us to use image transformations to control corresponding music transformations.

3. VIS2MUS: THE USER INTERFACE

Based on the visual-to-music mapping, we develop the Vis2Mus system as a controllable music generation interface, as shown in Figure 6. Users can generate music by following the steps below:

- 1. Select initial melody-sketch and polyphony-image anchor pairs from drop-down boxes in ⑤. The selected items (which are prelabeled) will be displayed in ① and ②;
- 2. **Conduct style transfer** as in Section 2.3. Pressing button (and the system will generate a fused image (displayed in (3)) as a visualization and preview of the fused music piece (displayed in (4)).
- 3. (Optional) Tuning brightness and contrast of generated music as in Section 2.1 and 2.2. Changing the brightness and contrast of the style-transferred image in ⑦ to further control the generated music in ④.



Fig. 6. A display of the interface of the Vis2Mus system.

4. CONCLUSION AND FUTURE WORK

In this study, we have contributed an analysis-by-synthesis method, which combines deep generative modeling and user study to explore interpretable visual-to-music mapping in a data efficient way. We discovered that visual-to-music mapping in the latent space has a nice property analogous to equivariant with respect to three transformations: changing brightness, changing contrast, and style transfer. The followup user study shows that the results are significant higher than random guess baselines. Inspired by the discoveries, we have also developed Vis2Mus, a controllable music generation interface using images and image transformations as the control handlers.

We are well-awared that we have just explored a small portion of visual-to-music mapping. In the future, we plan to study more forms of cross-modal transformation acting on more possible latent factors. We also plan to conduct a much larger scale user study, from which more interesting multimodal patterns can potentially be discovered.

5. REFERENCES

- [1] Wenzhong Guo, Jianwen Wang, and Shiping Wang, "Deep multimodal representation learning: A survey," *IEEE Access*, vol. 7, pp. 63373–63394, 2019.
- [2] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al., "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning*. PMLR, 2021, pp. 8748–8763.
- [3] Zihao Wang, Xihui Liu, Hongsheng Li, Lu Sheng, Junjie Yan, Xiaogang Wang, and Jing Shao, "CAMP: Cross-modal adaptive message passing for text-image retrieval," in *Proceedings of the IEEE/CVF Interna*tional Conference on Computer Vision, 2019, pp. 5764– 5773.
- [4] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever, "Zero-shot text-to-image generation," in *International Conference on Machine Learning*. PMLR, 2021, pp. 8821–8831.
- [5] MD Zakir Hossain, Ferdous Sohel, Mohd Fairuz Shiratuddin, and Hamid Laga, "A comprehensive survey of deep learning for image captioning," *ACM Computing Surveys (CsUR)*, vol. 51, no. 6, pp. 1–36, 2019.
- [6] Xuanchi Ren, Haoran Li, Zijian Huang, and Qifeng Chen, "Self-supervised dance video synthesis conditioned on music," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 46–54.
- [7] Kun Chen, Yusong Wu, Ziyue Wang, Xuan Zhang, Fudong Nian, Shengchen Li, and Xi Shao, "Audio captioning based on transformer and pretrained cnn," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events Workshop*, 2020, pp. 21–25.
- [8] Rafael Valle, Kevin Shih, Ryan Prenger, and Bryan Catanzaro, "Flowtron: an autoregressive flow-based generative network for text-to-speech synthesis," *arXiv* preprint arXiv:2005.05957, 2020.
- [9] Ilaria Manco, Emmanouil Benetos, Elio Quinton, and György Fazekas, "MusCaps: Generating captions for music audio," in 2021 International Joint Conference on Neural Networks (IJCNN). IEEE, 2021, pp. 1–8.
- [10] Yixiao Zhang, Ziyu Wang, Dingsu Wang, and Gus Xia, "BUTTER: A representation learning framework for bidirectional music-sentence retrieval and generation," in *Proceedings of the 1st workshop on nlp for music and audio (nlp4musa)*, 2020, pp. 54–58.

- [11] Yixiao Zhang, Junyan Jiang, Gus Xia, and Simon Dixon, "Interpreting song lyrics with an audio-informed pre-trained language model," *arXiv preprint arXiv*:2208.11671, 2022.
- [12] Xiaodong Tan, Mathis Antony, and H Kong, "Automated music generation for visual art through emotion.," in *ICCC*, 2020, pp. 247–250.
- [13] Elena Rivas Ruzafa, *Pix2Pitch: generating music from paintings by using conditionals GANs*, Ph.D. thesis, ETSI_Informatica, 2020.
- [14] Stan Z. Li and Anil Jain, Eds., *Analysis-by-Synthesis*, pp. 35–36, Springer US, Boston, MA, 2009.
- [15] Wassily Kandinsky, *Concerning the spiritual in art*, Courier Corporation, 2012.
- [16] Ziyu Wang, Dingsu Wang, Yixiao Zhang, and Gus Xia, "Learning interpretable representation for controllable polyphonic music generation," *arXiv preprint arXiv:2008.07122*, 2020.
- [17] Ziyu Wang, Yiyi Zhang, Yixiao Zhang, Junyan Jiang, Ruihan Yang, Junbo Zhao, and Gus Xia, "PianoTree VAE: Structured representation learning for polyphonic music," *arXiv preprint arXiv:2008.07118*, 2020.
- [18] Yongcheng Jing, Yezhou Yang, Zunlei Feng, Jingwen Ye, Yizhou Yu, and Mingli Song, "Neural style transfer: A review," *IEEE transactions on visualization and computer graphics*, vol. 26, no. 11, pp. 3365–3385, 2019.
- [19] Hang Zhang and Kristin Dana, "Multi-style generative network for real-time transfer," *arXiv preprint arXiv:1703.06953*, 2017.
- [20] Jingwei Zhao and Gus Xia, "AccoMontage: Accompaniment arrangement via phrase selection and style transfer," *arXiv preprint arXiv:2108.11213*, 2021.