Retaining Semantics in Image to Music Conversion

Zeyu Xiong
The Hong Kong University of Science and Technology (Guangzhou)
Guangzhou, China
zxiong666@connect.hkust-gz.edu.cn

Pei-Chun Lin
Feng Chia University
Taichung City, Taiwan
peiclin@fcu.edu.tw

Amin Farjudian
University of Nottingham Ningbo China
Ningbo, China
Amin.Farjudian@gmail.com

***Abstract*—We propose a method for generating music from a given image through three stages of translation, from image to caption, caption to lyrics, and lyrics to instrumental music, which forms the content to be combined with a given style. We train our proposed model, which we call BGT (BLIP-GPT2-TeleMelody), on two open-source datasets, one containing over 200,000 labeled images, and another containing more than 175,000 MIDI music files. In contrast with pixel level translation, the BGT model retains the semantics of the input image. We verify our claim through a user study in which participants were asked to match input images with generated music without access to the intermediate caption and lyrics. The results show that, while the matching rate among participants with low music expertise is essentially random, the rate among those with composition experience is significantly high, which strongly indicates that some semantic content of the input image is retained in the generated music.**

***Index Terms*—media semantics, media composition, machine learning**

I. Introduction

With the growth of computational power, simulation of human artistic creation by artificial intelligence (AI) has gained further prominence, with advances which not only provide us with further insight into the nature of artistic creation, but also have ramifications for creative industries [1]. In AI composition, musicality of the generated piece has garnered much of the attention of the researchers, and is still a significant challenge.

In this article, we focus on another salient feature, i. e., semantics. We consider the visual medium of images and the audio medium of music. We propose a model which transforms an input image to an instrumental music track via the intermediate steps of image caption generation, lyrics generation from the image caption, and finally lyrics to music translation. The components used for this purpose are the image captioning method of Bootstrapping Language-Image Pre-training (BLIP) [2], a lyrics generator based on Generative Pre-trained Transformer 2 (GPT-2) [3], and the lyrics to music translator TeleMelody [4]. As such, we use the initials of the three components and refer to the pipelined system as BGT.

We train BGT on two open-source datasets: the COCO Dataset [5], which contains over 200,000 labeled images, and Lakh MIDI Dataset [6], which contains over 175,000 MIDI music files. We augment BGT with the one-shot music style migration system Groove2Groove [7], to incorporate various genres of music. A schema of the system is depicted in Fig. 1 and the code is available at:

- https://github.com/BILLXZY1215/BGT-G2G

While images tend to have tangible content, the debate over whether music admits semantics has not been settled. Based on Gricean criteria, Mikalonytė [8] argues that pure music does not have semantic content. Jackendoff [9] states that music is not propositional and it cannot convey messages about people or objects. On the other hand, in a series of articles, Schlenker has recently developed a semantic framework for music (see [10] and the references therein).

Our aim is not to address whether music admits a Tarski-style semantics as is applicable to natural or formal languages. The paradigm that we adopt is closer to that of Patel [11], according to which, as long as a musical element brings to mind things other than itself, it indicates semantic content. Thus, in our evaluations, we take human usage and perception as the primary indicators of semantics and its retention [12]. For the evaluation of the model, we have carried out a user study to investigate the extent of semantics retention of the system from the perspective of individuals with various levels of musical expertise. The participants were asked to match a set of images with the generated music tracks without access to the intermediate captions and lyrics. The results show that the participants with higher musical expertise (especially those with composition experience) have a significantly higher success rate at matching images with their corresponding music tracks, compared with participants without music expertise.

II. Background and Related Work

The field of AI music composition has a rich literature, and may be regarded as a microcosm of modern machine learning (ML). For instance, among the variety of ML models used in composition, we mention LSTM [13], [14], auto encoder [15], RBM [16] and GAN [17]. Acceptable levels of musicality have been achieved in projects such as MidiNet [18], MuseGAN [19] and Jukebox [20].

While these systems perform well in arranging notes and chords, the emotions that they evoke cannot be compared against a tangible reference. Visual perception, on the other hand, can play a significant role in composing and evaluating

This research was supported by the Ministry of Education, R.O.C., under the grant TEEP@AsiaPlus, the Ministry of Science and Technology, R.O.C., under the grant No. MOST 109-2221-E-035-063-MY2, and by Feng Chia University under the 2022 Project Research Grant.

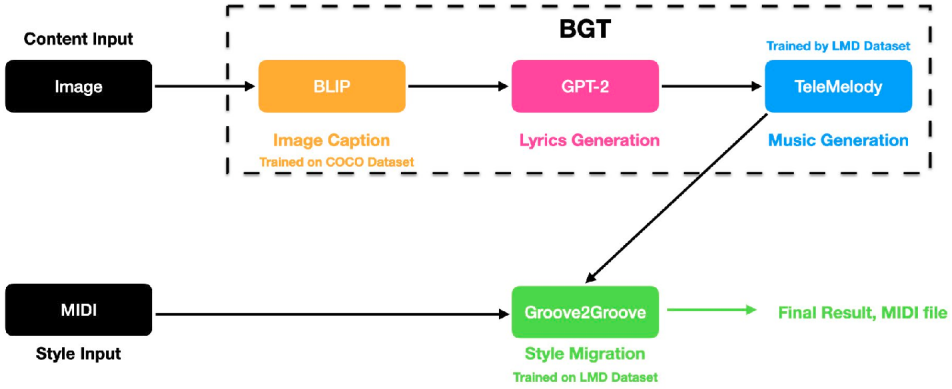

Fig. 1. BGT-Groove2Groove: Pipeline.

music [21]–[23]. Hence, we focus on some related work on image to music conversion.

Music can be represented in audio format (e. g., WAVE, Flac, MP3, etc.) or symbolic format (e. g., MIDI, Piano roll matrix, etc.). Some direct methods of image to music conversion regard an image as a matrix of pixels, and generate a piano roll from the matrix. For instance, according to Scriabin's concept of synesthesia [24], the twelve notes of the diatonic scale can be mapped to twelve different colors in the visible spectrum, which can be the basis for direct conversion of images to music. Other direct methods include methods of Mathigatti[1] and Vooydzig[2].

There are non-direct methods of image to music conversion as well, such as Musical Vision [25] and EdgeSonic [26], where the focus is on the time series generated according to how human eye scans various parts of an image. Other notable methods include image sonification [27]. Nevertheless, these methods are also incapable of retaining the semantic information of an input image.

The relationship between pictorial and music semantics has been studied before. For instance, Schlenker [10] has developed formal semantics for both within the framework of Super Semantics, although his focus has been mostly on sequences of images. Wu et al. [28], [29] have also devised algorithmic methods for estimating the semantic match between images and music. The experiments in [29] were performed over songs that contain lyrics. In contrast, our generated music tracks are instrumental. As such, they contain no verbal indicators of any extra-musical semantic clues.

A more important distinguishing feature of our work is that the music tracks are generated entirely by machine learning systems, while in [28], [29], the music is created by musicians. It should be emphasized that we make no claims as to whether the current machine learning systems in general—and those used in our model in particular—exhibit intelligence or understanding comparable to human beings. In fact, we are very skeptical of such claims. Nonetheless, we hope that our work sheds some light on the capabilities of such systems in retention of semantics in cross-form art creation, and also on our understanding of human perception of different forms of art.

[1]https://github.com/mathigatti/midi2img
[2]https://github.com/vooydzig/img2midi

III. BGT Pipeline

In this section, we present the BGT pipeline in more details. Our focus is the image content, as opposed to (say) coloring or style. To that end, the first step involves generation of caption from a given image. There are many image captioning methods available based on, e. g., LSTM [30], [31], GAN [32], or attention mechanism [2], [33]. We have chosen Bootstrapping Language-Image Pre-training (BLIP) [2] as the state-of-the-art component based on vision-language pre-training (VLP). For example, when given the album cover of Abbey Road by The Beatles (Fig. 2), BLIP generates the caption *"A group of people walking across the road"*.

BLIP incorporates a multimodal hybrid encoder-decoder (MED)—which can be used for various tasks—to pre-train a unified model with comprehension and creation capabilities. A unimodal encoder, an image-grounded text encoder, and an image-grounded text decoder are integrated into BLIP to achieve better performance.

To bring the content closer to a musical form, we transform the caption into lyrics using OpenAI's Generative Pre-trained Transformer 2 (GPT-2) [3]. The first few verses of the lyrics generated for Abbey Road album cover are as follows:

> Well, I'm a rolling stone
> But my song is never sung
> I got no mob traditions
> Got 'em rocking and a-rollin'
> ...

For the final stage of the pipeline, we have used TeleMelody [4] to generate music from lyrics. TeleMelody performs the translation in two stages: lyrics-to-template and

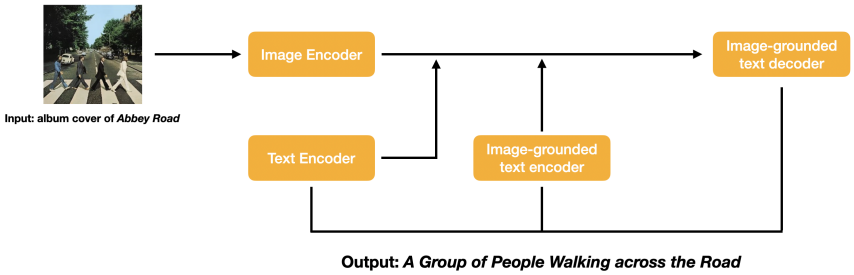

Fig. 2. A working example of BLIP: Image Captioning

template-to-melody. Musical templates for tonality, chord progressions, rhythmic patterns, and tempo are used as intermediate steps to enable smoother transition from lyrics to melody.

Generating a template from lyrics requires two simultaneous steps:

1) Lyrics to rhythm mapping, with a default tonality.
2) Punctuation to cadence mapping.

Punctuation is essential for chord changes and melodic variations. The pipeline, and the dependencies of various components are depicted in Fig. 3, where it can be seen that, for generating melody from a template:

- the pitch depends on tonality, chord progression, and cadence;
- position depends on the rhythm pattern;
- duration depends on cadence.

While Telemelody can perform automatic segmentation of lyrics and chord matching, it can also accept the chord sequence as a parameter to customize the output.

A. Training and preparation

As illustrated in Fig. 1, BGT has three components:

1) **Image to caption:** For image captioning task, we use the COCO Dataset [5]—with over 200,000 labeled images—to train the BLIP model.
2) **Caption to lyrics:** GPT-2 has been trained by the WebText Dataset [3], which is an internal OpenAI corpus created by scraping web pages with emphasis on document quality. Thus, we directly use Mathigatti's pre-trained model.[3]
3) **Lyrics to music:** To convert lyrics to music, we train TeleMelody using the Lakh MIDI Dataset [6], which contains more than 175,000 MIDI files, ranging over various music styles. We use Lakh MIDI Dataset for both lyrics-to-template and template-to-melody components of TeleMelody.

We enrich the output of BGT by augmenting the model with Groove2Groove [7], a one-shot AI system for style transfer, which is implemented using an encoder-decoder model.

[3]https://lyrics.mathigatti.com/

IV. Experiments

We take four images which evoke distinct emotions, and use these images as input to the pipeline of Fig. 1, resulting in four instrumental music tracks. Then, we ask the survey participants to match the four tracks generated with the four images. There are $4! = 24$ possible ways of matching the tracks with images. Thus, if the matching is done randomly, there is around $1/24 \approx 4\%$ chance of success.

After gathering the results of our user study, if the matching rate is close to random (i. e., 4%) then this will mean that not much of the semantic content of the input images has been retained through the pipeline. If the matching success rates are significantly higher than random, then we have some evidence that the pipeline indeed retains some of the semantic content of the input.

The four images that we have selected are shown in Fig. 4. The titles for these images are, from left to right: Mysterious Man, Coffee & Book, Lovers in the Rain, and Big Wave. The emotions commonly attributed to these images by our participants are also shown underneath each image in Fig. 4.

For the style input of Groove2Groove, we select four songs from Lakh MIDI Dataset [6] that are representative of the four styles of blues, country, electronic, and jazz music. This results in a total of 16 combinations of input images and music genres.

We also assign a chord progression to each intermediate lyrics, which forms the input to TeleMelody. We have chosen 16 chord progressions (with a one-to-one match with the 16 combinations mentioned above) to increase randomness. The participants are not informed about these chord progressions.

On average, when deployed using the trained models, the time required to generate a music track from the input image and the style is around five minutes.

V. Evaluation

The quality of the output of the pipeline, and how well the output matches the input, are ultimately best judged by human participants. Hence, we conducted a user study in which the participants were asked to rate various aspects of the input and output in terms of emotions that they evoke and how they match. For the experiments, the participants were given a list of ten emotion tags (psychological attributes) as follows: intense, calming, romantic, thrilling, mellow, manic, happy,

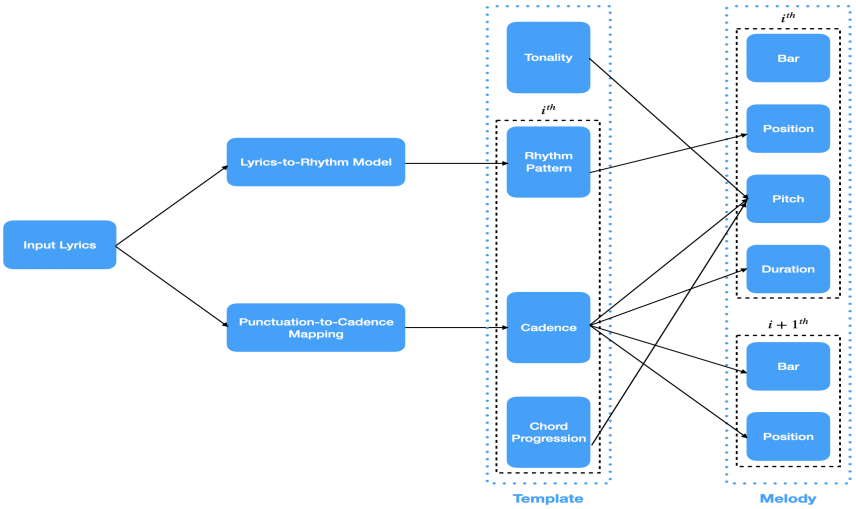

Fig. 3. TeleMelody: Pipeline.

Fig. 4. Input images of the pipeline: Four images that evoke distinct emotions.

sad, deep and dreamy. These were chosen from 38 attributes listed in [34, Table 1]. We selected only ten attributes for convenience of the participants, which we assumed would suffice for the four input images.

The questionnaire was anonymized, and the participants were informed that their response would be used only for academic research. Furthermore, the participants were informed that the output music tracks were computer generated based on the input images.

In the design of the questionnaire and the sequencing of questions, we took inspiration from previous work on music generation which involved user study, e. g., music quality marking [35], [36] and emotion marking [37]. On average, each participant had to spend around fifteen minutes on the survey.

Remark 5.1: We estimated that fifteen minutes was an acceptable duration for each participant to help us with the user study. Thus, the main reason behind choosing four images and generating four music tracks—as opposed to a smaller or larger number—was to strike the right balance between reliability of the statistical analyses and convenience of the participants.

The participants were asked to respond to the following requests in the survey:

1) Select music expertise level: nonexpert, familiar with basic music theory, and having experience in composing music.
2) For each of the given images (Fig. 4), select a best matching emotion tag (psychological attribute) from the list of the ten tags: intense, calming, romantic, thrilling, mellow, manic, happy, sad, deep and dreamy.
3) Select the music style (genre) of your choice: The participants were given the chance to listen to the style input tracks from each of the four genres of blues, country, electronic, and jazz. Then, they were asked to choose one from the four.
4) Evaluate the result: At this stage, the participants were given the generated tracks, and they were asked to rate them based on three aspects, on the scale of 1 to 5:
   - *Similarity* between the generated track and the style input (very low similarity 1 – 5 very high similarity).
   - *Musicality* of the result (not musical 1 – 5 satisfactory).
   - *Creativity* of the model (non-creative 1 – 5 very creative).
5) Select the best matching emotion tag (out of ten) for each generated track.
6) Match each of the four generated tracks with the corresponding input image: This is the most important step

TABLE I
DISTRIBUTION BY GENRE PREFERENCE.

Genre	Percentage
Blues	24.67%
Country	24.67 %
Electronic	25.33%
Jazz	25.33 %

TABLE II
DISTRIBUTION BY EXPERTISE LEVEL.

Expertise Level	Percentage
Nonexpert	49.33%
Basic Music Theory	24.00 %
Experience in Composing	26.67%

in the survey, which determines whether there is any semantic correspondence between the input images and the generated tracks, in a way that is recognizable by participants.

The survey was taken by a total of 150 participants with different expertise levels and genre preferences. As shown in Table I, in terms of genre preference, there was an almost perfect even distribution. In terms of expertise level, as shown in Table II, almost half of the participants were nonexperts, while those who had familiarity with basic music theory and those who had experience in composition made up about a quarter of the participants each.

A. Matching Emotion Tags

In steps 2 and 5 of the survey, the participants were asked to assign an emotion tag to each image and the generated tracks. When choosing the tag for the generated track, if the selection is done randomly (out of the ten tags) then the correct matching rate should be 10%. As can be seen from Fig. 5:

1) The matching rate increases with expertise in music among the participants.
2) In some cases, the matching rate of nonexperts is not significantly different from the random 10%, e. g., for the input images 'Lovers in the Rain' and 'Mysterious Man'.
3) The overall matching rate is generally significantly higher than random.

B. Similarity, Musicality and Creativity

In Step 4 of the survey, the participants were asked to rate the output in terms of the three features of similarity (with the input style track), musicality of the generated track, and creativity in composition exhibited by the model. Fig. 6 shows the mean value of the responses provided by the participants, broken down by genre.

As can be seen, the participants gave a high score for similarity across the four genres. The lowest mean score was given as 3.784. There are more noticeable variations in scores for musicality and creativity. The genre of Jazz was given the highest average score for musicality (and also the highest score in the subsequent image-music matching, as shown in Fig. 8). Country music received the highest average score for creativity, while the scores were the lowest for blues music, both for musicality and creativity.

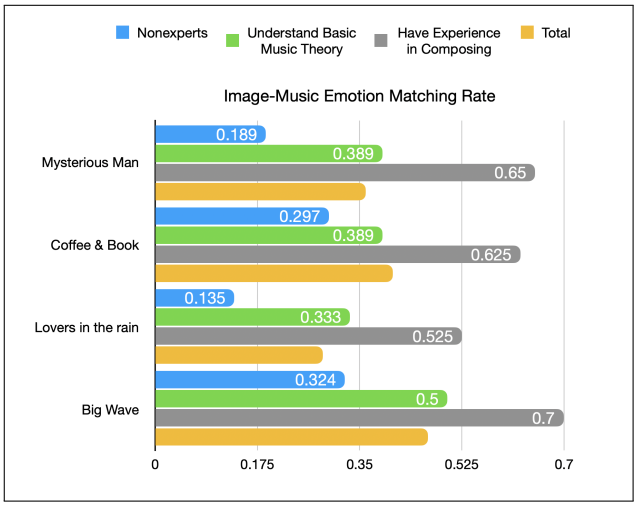

Fig. 5. Image-music emotion matching rate.

TABLE III
SIGNIFICANCE TEST OF DIFFERENCES IN USER'S MEAN SCORES.

Hypothesis	P-value	Significance
Similarity: Blues vs Country	0.056	Not Significant
Musicality: Jazz vs Blues	2.26×10^{-18}	Highly Significant
Creativity: Blues vs Country	2.18×10^{-10}	Highly Significant

At first sight, the scores reported in Fig. 6 seem quite close to one another. Thus, we investigated the significance of the differences in the scores by performing significance test between the two groups with the largest differences in terms of similarity, musicality, and creativity. For example, in similarity, Blues (3.993) and Country (3.784) are the two genres with the largest differences.

We apply *Analysis of Variance* (ANOVA) test to ascertain significance of differences. The basis for our judgement is the following inequalities regarding the *P-Value*:

$$\begin{cases} \text{P-value} \geq 0.05 \implies \text{Not Significant}, \\ \text{P-value} < 0.05 \implies \text{Significant}, \\ \text{P-value} < 0.01 \implies \text{Highly Significant}. \end{cases}$$

Table III shows the *P-Value* between the scores with the largest differences, for the three aspects of similarity, musicality, and creativity, and their corresponding significance. We conclude that, for musicality and creativity, the differences in participants' mean scores are significant, while for similarity, the differences are not significant. This may be interpreted as an evidence for consistency of the model in terms of musical style retention.

In summary, the model performed well in terms of musical style retention. The performance, however, was not uniform for musicality and creativity across genres. We leave an investigation of the reasons behind this for future work.

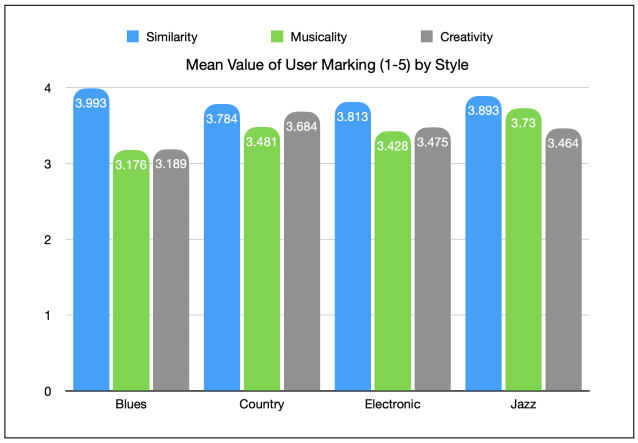

Fig. 6. Ratings of similarity, musicality, and creativity.

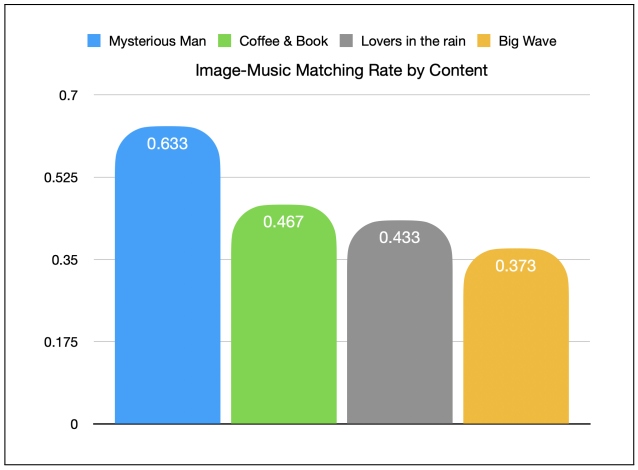

Fig. 7. Image-music matching rate by image content.

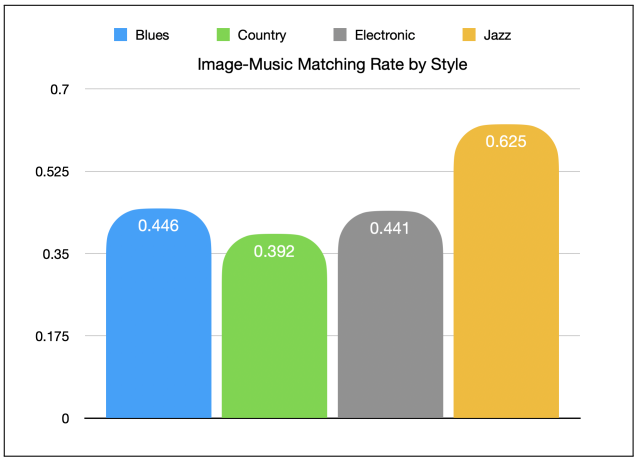

Fig. 8. Image-music matching rate by genre (style track).

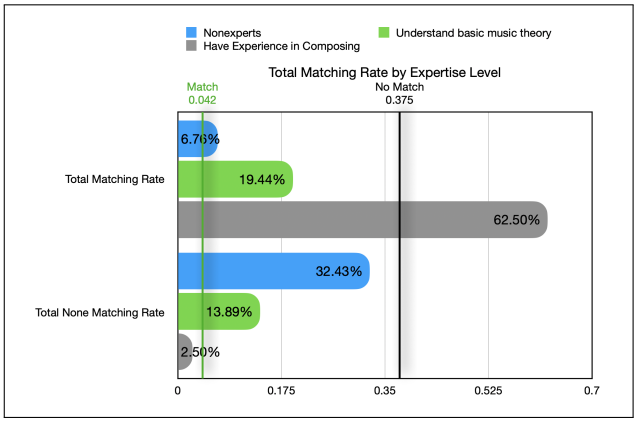

Fig. 9. Overall matching rates.

C. Image-Music Matching

Step 6 of the survey contains, arguably, the most important question, where the participants are asked to match the output music tracks with the input images. The participants do not have access to the intermediate captions and lyrics, and do not know the architecture of the model used in the generation of the tracks. The image-music matching rate by image content can be seen in Fig. 7, which shows that 'Mysterious Man' has had the highest correct matching rate, whereas 'Big Wave' scores the lowest.

The image-music matching rate by genre is shown in Fig. 8, according to which, Jazz genre received the highest matching score, and Country scores the lowest.

Finally, we consider the overall matching rates, broken down by the expertise of the participants, as shown in Fig. 9.

As there are four input images and four output tracks, there are $4! = 24$ ways of matching the output music to the input images. If done randomly, the success rate should be around $1/24 \approx 0.042$, and total failure (i. e., not matching any output music to the corresponding input image correctly) would have the probability $9/24 = 0.375$. This is because the number of the so-called *derangements* of a set of 4 elements is 9 [38].

As can be seen from Fig. 9:

1) Correct matching rates increase with the participants' expertise in music. For nonexperts, the matching rate is not significantly different from random match, while the rate for participants with composition experience is quite stark, at 62.50%, which is well above the 4.2% line.
2) The rate of totally incorrect matching (i. e., derangement) among nonexperts (32.43%) is also not far from the 37.5% random threshold, whereas, for those with composition experience, the rate of derangement is only 2.50%, which is far below the random threshold.

VI. Discussion

One might expect skilled musicians to perceive links between music *created by musicians* and other forms of art at a rate higher than nonexperts. Yet, the very fact that, in our experiments, the matching rates by music experts are so higher than random matching, prompts further analysis and explanation, given that all the components of the BGT-Groove2Groove pipeline of Fig. 1 are *machine learning* models. This is in contrast with (say) the experiments of [28], [29] which are performed on music created by musicians, with music videos

(also created by artists) that have been designed carefully to match the music.

We believe that it is very unlikely that any of the machine learning components of our model in isolation, or combined, can exhibit human intelligence, let alone understanding of human emotions. It is, rather, more likely that retention of semantics and emotions is achieved because of the nature of datasets used in training of the components. For instance, GPT-2 is a transformer which operates based on the attention mechanism. In other words, it is designed to 'learn' the correlations among various items in an input sequence, and trained on a very large dataset. Thus, it ultimately operates based on pattern matching.

The datasets for all the components of the pipeline are prepared by human beings. As such, it is possible that there are strong patterns common to images, text, and music, that are discovered by the machine learning components, which are subsequently manifested in the output of the system. We stress that the image content that is passed through the pipeline is not 'syntactical', e. g., not pixel level color distribution. Therefore, the patterns that appear in the output music must be more related to the intermediate text (i. e., caption and lyrics) rather than the color distribution of the input image.

Mehr et al. [39] have demonstrated that music appears in every society observed, and identifiable acoustic features of songs (e. g., accent, tempo, pitch range, etc.) predict their primary behavioral context (love, healing, etc.). Thus, one may conclude that people (of all societies) have some hard-coded familiarity with the context of music that they hear. This, in turn, may be reflected in the datasets used to train our systems.

VII. Concluding remarks

We have proposed a model, which we call BGT, for generating music from given input images. The focus is on the semantic content of the input image, as opposed to (say) color distribution. Thus, the model, depicted in Fig. 1, works based on the three stages of image to caption, caption to lyrics, and lyrics to (instrumental) music transformation. All the stages are carried out by machine learning models.

Through a user study with 150 participants, we have demonstrated that, with enough musical expertise, the participants can match the output music with the input image, and also assign consistent emotional tags to both. The results of this kind may be helpful in investigation of (limits of) intelligence and understanding exhibited by machine learning models. The approach can also be helpful for similar work in cross-media conversion not restricted to image and music.

In practical terms, such models can be useful for automatic composition of short music pieces that are relevant to specific contexts, e. g., composition of background music for advertisement videos.

As for future work, we envisage some directions as follows:

1) A convincing theoretical investigation of why the matching rates can be significantly higher than random.
2) Improvements on the model structure and datasets to enable generation of higher quality music with more diversity.
3) Improvements on the control of the experiment conditions. For instance, including control data such as non-related images and music to the image and music set will strengthen the conclusions of the current work.
4) Investigation of the influence of genre on the qualitative features of the generated music. We have seen in Fig. 6 that the scores for musicality and creativity are different across various genres. Whether this is due to some intrinsic features of each genre, or deficiencies of the datasets and training, will be investigated in future work. Recall that the style input tracks in the current work have been limited to four representative tracks from each genre. A broader selection of style tracks from each genre will be considered in future work.
5) Currently, we have fixed the four input images and the accompanying styles. It will be helpful to automate the entire pipeline for any input image and style, and survey a broader base of participants. This will shed more light on the potentials and limitations of such approach to music composition.
6) Carrying out the experiment in languages other than English. For the current work, the intermediate steps have involved English text. If the results of the current study are sound and robust, then, in light of the results of [39], similar patterns are likely to exist in languages other than English, which link semantics of images to the patterns that are generated in the resulting music.
7) In the opposite direction, design of a method for generating images—with concrete content—from pure music, in such a way that enables a high rate of matching, at least by music experts.

   In this respect, we point out the related work of Passalis and Doropoulos [40], who have used deep learning for translating music to sequences of images (visual stories) with the aim of reflecting the sentiment of the input music track. While we have used caption and lyrics in the intermediate steps, in [40], the authors use valence and arousal [41]. Furthermore, their aim is retaining sentiment in generation of visual stories, while we will focus on retention of semantics in generation of a single image.

References

[1] N. Anantrasirichai and D. Bull, "Artificial intelligence in the creative industries: a review," *Artificial Intelligence Review*, vol. 55, pp. 589–656, 2022.

[2] J. Li, D. Li, C. Xiong, and S. C. H. Hoi, "BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation," *CoRR*, vol. abs/2201.12086, 2022. [Online]. Available: https://arxiv.org/abs/2201.12086

[3] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever *et al.*, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.

[4] Z. Ju, P. Lu, X. Tan, R. Wang, C. Zhang, S. Wu, K. Zhang, X. Li, T. Qin, and T. Liu, "Telemelody: Lyric-to-melody generation with a

template-based two-stage method," *CoRR*, vol. abs/2109.09617, 2021. [Online]. Available: https://arxiv.org/abs/2109.09617

[5] T. Lin, M. Maire, S. J. Belongie, L. D. Bourdev, R. B. Girshick, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: common objects in context," *CoRR*, vol. abs/1405.0312, 2014. [Online]. Available: http://arxiv.org/abs/1405.0312

[6] C. Raffel, "Learning-based methods for comparing sequences, with applications to audio-to-midi alignment and matching," Ph.D. dissertation, Columbia University, 2016.

[7] O. Cífka, U. Şimşekli, and G. Richard, "Groove2groove: One-shot music style transfer with supervision from synthetic data," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2638–2650, 2020.

[8] E. S. Mikalonytė, "Why does pure music not have semantic content?" *Revista Portuguesa de Filosofia*, vol. 74, no. 4, pp. 1355–1376, 2018.

[9] R. Jackendoff, "Parallels and nonparallels between language and music," *Music Perception: An Interdisciplinary Journal*, vol. 26, no. 3, pp. 195–204, 2009.

[10] P. Schlenker, "Musical meaning within super semantics," *Linguistics and Philosophy*, vol. 45, pp. 795–872, 2022.

[11] A. Patel, *Music, Language, and the Brain*. Oxford: Oxford University Press, 2010.

[12] P. Saint-Dizier, *Music and Artificial Intelligence*. Cham: Springer International Publishing, 2020, pp. 503–529.

[13] D. Eck and J. Schmidhuber, "A first look at music composition using LSTM recurrent neural networks," Istituto Dalle Molle Di Studi Sull Intelligenza Artificiale, Tech. Rep., 2002.

[14] J. Wu, C. Hu, Y. Wang, X. Hu, and J. Zhu, "A hierarchical recurrent neural network for symbolic melody generation," *IEEE transactions on cybernetics*, vol. 50, no. 6, pp. 2749–2757, 2019.

[15] M. Bretan, G. Weinberg, and L. P. Heck, "A unit selection methodology for music generation using deep neural networks," *CoRR*, vol. abs/1612.03789, 2016. [Online]. Available: http://arxiv.org/abs/1612.03789

[16] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *science*, vol. 313, no. 5786, pp. 504–507, 2006.

[17] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.

[18] L. Yang, S. Chou, and Y. Yang, "MidiNet: A convolutional generative adversarial network for symbolic-domain music generation using 1D and 2D conditions," *CoRR*, vol. abs/1703.10847, 2017. [Online]. Available: http://arxiv.org/abs/1703.10847

[19] H.-W. Dong, W.-Y. Hsiao, L.-C. Yang, and Y.-H. Yang, "MuseGAN: Multi-track sequential generative adversarial networks for symbolic music generation and accompaniment," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.

[20] P. Dhariwal, H. Jun, C. Payne, J. W. Kim, A. Radford, and I. Sutskever, "Jukebox: A generative model for music," *CoRR*, vol. abs/2005.00341, 2020. [Online]. Available: http://arxiv.org/abs/2005.00341

[21] C.-J. Tsay, "Sight over sound in the judgment of music performance," *Proceedings of the National Academy of Sciences*, vol. 110, no. 36, pp. 14 580–14 585, 2013.

[22] S.-L. Tan, A. J. Cohen, S. D. Lipscomb, and R. A. Kendall, *The psychology of music in multimedia*. Oxford University Press, 2013.

[23] F. Bravo, "The influence of music on the emotional interpretation of visual contexts," in *International Symposium on Computer Music Modeling and Retrieval*. Springer, 2012, pp. 366–377.

[24] K. Peacock, "Synesthetic perception: Alexander scriabin's color hearing," *Music perception*, vol. 2, no. 4, pp. 483–505, 1985.

[25] A. Polo and X. Sevillano, "Musical vision: An interactive bio-inspired sonification tool to convert images into music," *Journal on Multimodal User Interfaces*, vol. 13, no. 3, pp. 231–243, 2019.

[26] T. Yoshida, K. M. Kitani, H. Koike, S. Belongie, and K. Schlei, "EdgeSonic: Image feature sonification for the visually impaired," in *Proceedings of the 2nd Augmented Human International Conference*, ser. AH '11. New York, NY, USA: Association for Computing Machinery, 2011. [Online]. Available: https://doi.org/10.1145/1959826.1959837

[27] W. S. Yeo and J. Berger, "A framework for designing image sonification methods," in *Proceedings of international conference on auditory display*, 2005.

[28] X. Wu, Y. Qiao, X. Wang, , and X. Tang, "Cross matching of music and image," in *Proceedings of the 20th ACM international conference on Multimedia (MM '12)*. New York, NY, USA: Association for Computing Machinery, 2012, pp. 837–840.

[29] X. Wu, Y. Qiao, X. Wang, and X. Tang, "Bridging music and image via cross-modal ranking analysis," *IEEE Trans. Multimedia*, vol. 18, no. 7, pp. 1305–1318, 2016.

[30] J. Gao, S. Wang, S. Wang, S. Ma, and W. Gao, "Self-critical n-step training for image captioning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 6300–6308.

[31] Y. Zheng, Y. Li, and S. Wang, "Intention oriented image captions with guiding objects," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8395–8404.

[32] L. Guo, J. Liu, P. Yao, J. Li, and H. Lu, "MSCap: Multi-style image captioning with unpaired stylized text," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4204–4213.

[33] M. Cornia, L. Baraldi, and R. Cucchiara, "Show, control and tell: A framework for generating controllable and grounded captions," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8307–8316.

[34] D. M. Greenberg, M. Kosinski, D. J. Stillwell, B. L. Monteiro, D. J. Levitin, and P. J. Rentfrow, "The song is you: Preferences for musical attribute dimensions reflect personality," *Social Psychological and Personality Science*, vol. 7, no. 6, pp. 597–605, 2016.

[35] J. Luo, X. Yang, S. Ji, and J. Li, "MG-VAE: Deep Chinese folk songs generation with specific regional style," *CoRR*, vol. abs/1909.13287, 2019. [Online]. Available: http://arxiv.org/abs/1909.13287

[36] S. Walter, G. Mougeot, Y. Sun, L. Jiang, K.-M. Chao, and H. Cai, "MidiPGAN: A progressive GAN approach to MIDI generation," in *2021 IEEE 24th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*. IEEE, 2021, pp. 1166–1171.

[37] G. C. Sergio, R. Mallipeddi, J.-S. Kang, and M. Lee, "Generating music from an image," in *Proceedings of the 3rd International Conference on Human-Agent Interaction*, 2015, pp. 213–216.

[38] R. Stanley, *Enumerative Combinatorics*, 2nd ed. Cambridge University Press, 2012, vol. 1.

[39] S. A. Mehr, M. Singh, D. Knox, D. M. Ketter, D. Pickens-Jones, S. Atwood, C. Lucas, N. Jacoby, A. A. Egner, E. J. Hopkins, R. M. Howard, J. K. Hartshorne, M. V. Jennings, J. Simson, C. M. Bainbridge, S. Pinker, T. J. O'Donnell, M. M. Krasnow, and L. Glowacki, "Universality and diversity in human song," *Science*, vol. 366, no. 6468, p. eaax0868, 2019.

[40] N. Passalis and S. Doropoulos, "deepsing: Generating sentiment-aware visual stories using cross-modal music translation," *Expert Systems with Applications*, vol. 164, p. 114059, 2021.

[41] D. M. Greenberg, M. Kosinski, D. J. Stillwell, B. L. Monteiro, D. J. Levitin, and P. J. Rentfrow, "The song is you: Preferences for musical attribute dimensions reflect personality," *Social Psychological and Personality Science*, vol. 7, no. 6, pp. 597–605, 2016.