COLORING MUSIC: BRIDGING MUSIC AND COLOR PALETTES FOR GRAPHIC DESIGN

Takayuki Nakatsuka **Masahiro Hamasaki** **Masataka Goto**
National Institute of Advanced Industrial Science and Technology (AIST), Japan
{takayuki.nakatsuka, masahiro.hamasaki, m.goto}@aist.go.jp

ABSTRACT

This paper explores the relationship between music and the color palettes used for designing their corresponding music cover images, providing a comprehensive analysis that bridges auditory and visual expression. Our findings reveal a relationship between musical pieces and certain colors, suggesting that the color palettes used in cover image design are carefully selected to reflect the auditory experience. Building on these findings, we propose a framework that estimates appropriate color palettes for musical pieces to support selecting colors for cover image design. Using a large private dataset of 582,894 pairs of a musical piece and its corresponding cover image from various music genres, our framework leverages deep learning techniques to train our color palette estimator. We demonstrate the effectiveness of our proposed framework in graphic design by showcasing an application that generates cover images using the estimated color palettes from given musical pieces.

1. INTRODUCTION

In multimodal music understanding, both music and their corresponding music cover images play a crucial role. For instance, Oramas et al. successfully improved music genre classification accuracy by incorporating image features in addition to audio features [1]. In addition, Lībeks and Turnbull showed that cover images involve distinct features that can be used to predict music genre tags [2]. These studies suggested that a cover image embodies the essence of its corresponding music content, thereby establishing that analyzing these images yields a deeper understanding of the music. This study focuses on the colors used in cover images and analyzes their relationship with music.

The colors used in cover images tend to empirically reflect the characteristics of the corresponding music style. As illustrated in Fig. 1, different music genres display distinctive characteristics in the colors used in the cover images. As colors are closely linked to cultural contexts [3], emotions [4, 5], and the ability to attract visual attention [6], cover images contribute to the promotion of music content and enhance the overall music appreciation experience [7]. Therefore, this relationship between music and the colors used in cover images has been the subject of several studies [8–10]. However, these studies have mainly focused on genres, not on musical pieces.

Figure 1. Example results of Google Search with the text queries "{music genre} music album covers," where we used the music genres 'Death metal,' 'Country,' and 'Electronic.' Music cover images for each genre are characterized by the colors used in cover image design: dark colors for 'Death metal,' brownish colors for 'Country,' and vivid colors for 'Electronic.'

This paper first investigates the preferred colors for designing cover images across multiple genres in our preliminary study (Section 4) and further explores the relationship between musical pieces and the colors used in their corresponding cover images based on our proposed framework (Section 5). In this study, we focus on not only a representative color but also color palettes used in cover images because they play a crucial role in graphic design [11–13], shedding light on the deliberate selection process of colors that reflect the essence of the music content.

Based on our findings that a relationship exists between musical pieces and the colors used in their corresponding cover images, we propose a framework to estimate appropriate color palettes for musical pieces. The key technical

aspects of our framework are how to extract color palettes from cover images and how to estimate color palettes for musical pieces. For a color palette extraction method, we employ *data-driven color manifolds* [14], which are useful in arranging the colors as a color palette. For a color palette estimator, we train a deep neural network to estimate an appropriate color palette for each musical piece. In this training, we leverage a pretrained audio model (*contrastive language-audio pretraining (CLAP)* [15] or *AudioToken* [16]) as an audio feature extractor to extract a distinctive feature from each musical piece. This framework bridges musical pieces and their corresponding cover images using color palettes.

To demonstrate the effectiveness of our framework, we present an example application that generates cover images using the estimated color palettes from given musical pieces to support creating visually appealing cover images.

2. RELATED WORK

Several studies have investigated the relationship between music and color. Wells argued that there is a correlation between music and color based on the principle of complementarity [17]. Furthermore, Pesek et al. suggested that since music and emotions are closely related (e.g., [18, 19]), as well as emotions and colors (e.g., [4, 5]), there exists a relationship between music and color mediated by emotions [20]. However, these studies have only partially elucidated the relationship between music and color, as they analyzed this relationship using a limited number of colors. Therefore, in this study, we use the colors used in music cover images that embody a musical essence [1, 2] as the basis for our analysis.

In research exploring the colors used in cover images, previous studies have focused on specific genres (classical [8] and metal [9]). Seker [8] discovered that the colors used in cover images for classical music predominantly favor neutral colors. Friconnet [9] found that cover images for metal music tend to use darker colors than those of other genres, with a preference for black and orange [9]. Although these studies provide insights into the colors used in cover images of specific genres, no studies have explored which color values are preferred for specific musical pieces of various genres.

Additionally, color themes used in designing cover images have been studied [10]. Dorochowicz and Kostek [10] analyzed cover images across multiple genres with respect to basic color analysis rules such as seasonal colors (e.g., spring (warm and bright), summer (cool and soft), autumn (warm and soft), and winter (cool and bright)) and degrees of brightness (e.g., light, medium, and dark). While their findings provide valuable insights into the color characteristics of each genre, they focus on a limited number of color palettes based on the basic color analysis rules.

In this paper, we investigate the relationship between musical pieces and the color palettes used in their corresponding cover images and explore the application of this relationship in cover image design.

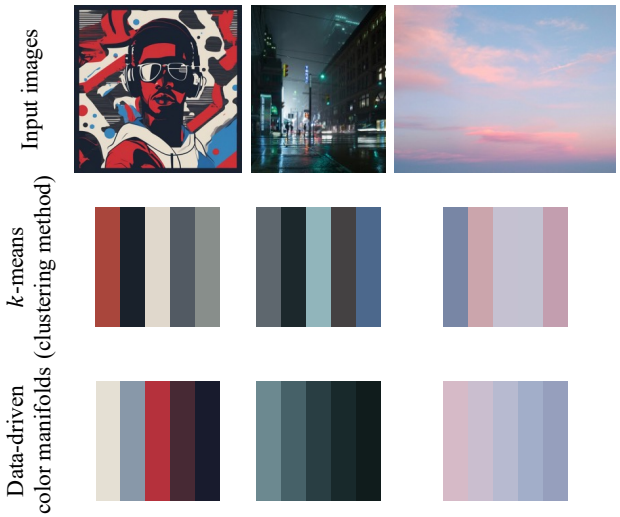

Figure 2. Comparison of color palette extraction methods. Given the input image (top row), the data-driven color manifolds (bottom row) extract a color palette from the image in consecutive color order, while k-means (middle row) extracts a color palette from the image in random color order.

3. COLOR EXTRACTION

To extract a representative color or color palettes from music cover images, we leverage *data-driven color manifolds* [14], a technique which aims to acquire color samples from images and learn a lower-dimensional manifold of the acquired color samples. The learned manifold reflects the distribution of colors in cover images, compressing areas of the color space that are less commonly used and expanding those that are more frequently utilized.

The technique involves several steps, starting with the acquisition of color samples from cover images. For successful color manifold learning, a sufficient number of samples (over 10k) must be obtained from each image. Note that we utilized all samples from $224\,\text{px} \times 224\,\text{px}$-resized cover images, amounting to over 50k samples. These samples are then used to estimate the density of each color in the cover images, with a focus on identifying and preserving the most important colors. A self-organizing map [21], which is used to reduce dimensionality, is then applied to derive the one-dimensional or two-dimensional color manifolds. We utilize the one-dimensional color manifold to extract color palettes from cover images. In practice, we calculate a discrete color manifold, which consists of $M \in \mathbb{N}$ colors, to use the derived color manifold as a color palette. All hyperparameter values related to density estimation and dimensionality reduction were taken from [14], except for the smoothness parameter, which we set to $r_0 = 1$.

The advantage of this technique over clustering methods such as k-means [22] is that the color palette extracted by the data-driven color manifolds has a meaningful ordering, where the order of colors is determined by the derived one-dimensional color manifold and thus results in consecutiveness, while the color palette extracted by a clustering method has a random ordering (see Fig. 2). When using a color palette consisting of multiple colors in graphic design,

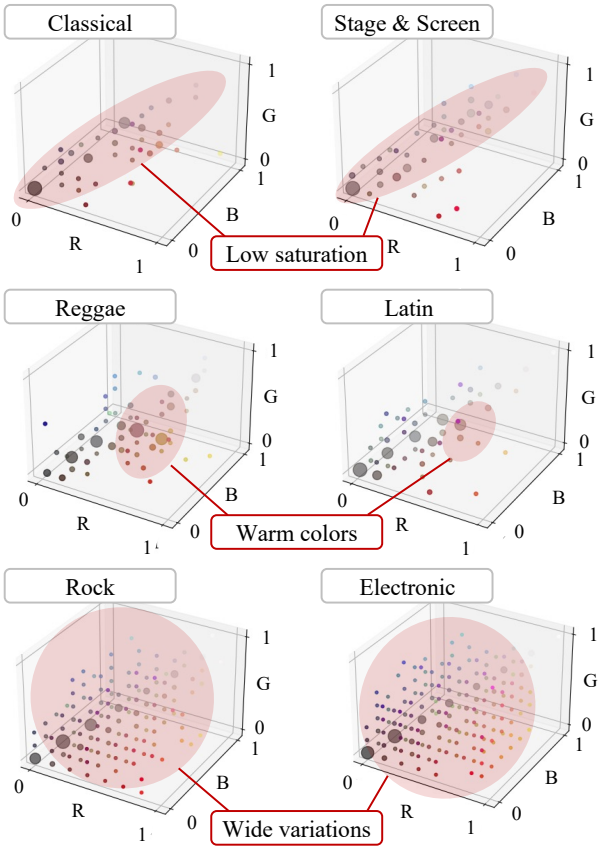

Figure 3. Visualization of a representative color used in music cover images by music genre. The larger the circle in the visualization, the more frequently the circle's color appears in the cover images.

the color palette with a continuous color order based on the data-driven color manifolds is intuitive and easy to use.

4. PRELIMINARY STUDY

This section describes our preliminary study that aims to analyze the preferred colors for designing music cover images across multiple genres by leveraging color palettes extracted from these images.

4.1 Experimental Setup

4.1.1 Dataset

We randomly collected 3,887 cover images (each image is an RGB image) for the experiments. We assigned genre tags to each image based on the grouping of genres and styles in Discogs [1], in which various music is organized into 15 genres and styles ('Blues,' 'Brass & Military,' 'Children's,' 'Classical,' 'Electronic,' 'Folk, World, & Country,' 'Funk / Soul,' 'Hip-Hop,' 'Jazz,' 'Latin,' 'Non-Music,' 'Pop,' 'Reggae,' 'Rock,' and 'Stage & Screen'). A total of 5,150 genre tags were assigned to 3,887 cover images, which means an average of 343.3 images per genre.

[1] The grouping of genres and styles in Discogs is available at https://support.discogs.com/hc/en-us/articles/360005055213-Database-Guidelines-9-Genres-Styles.

Table 1. List of representative colors most frequently used in music cover images for each music genre, excluding grayscale colors.

Music genre	RGB value	Color
Blues	(148,135,102)	
Brass & Military	(110,101,74)	
Children's	(139,178,241)	
Classical	(105,132,128)	
Electronic	(72,36,36)	
Folk, World, & Country	(108,101,68)	
Funk / Soul	(111,109,73)	
Hip-Hop	(73,36,36)	
Jazz	(181,145,109)	
Latin	(146,112,110)	
Non-Music	(165,127,156)	
Pop	(110,73,73)	
Reggae	(168,132,68)	
Rock	(72,36,36)	
Stage & Screen	(174,172,106)	

4.1.2 Implementation details

For representing colors in color manifolds, we utilized an RGB color space, which is a widely used additive color model. We resized all of the cover images into 224 px $\times$ 224 px and normalized their RGB values to [0, 1]. Then, for the purpose of this preliminary study, we simply extracted one representative color (i.e., $M = 1$) from the resized images using the data-driven color manifolds [14] as described in Section 3. Note that we extracted more colors to form the color palettes in Section 5.4. We used all of the pixels in the resized images as color samples. The constructed color manifold is divided into eight bins, and the samples are discretized into these bins, enabling their visualization as a histogram.

4.2 Results

Fig. 3 represents three-dimensional histograms for the R, G, and B values in the RGB color space. As shown in Fig. 3, trends in the distribution of colors differ by genres: for example, 'Classical' and 'Stage & Screen' music cover images have low saturation, and 'Reggae' and 'Latin' music cover images tend to use warm colors such as red and yellow. Additionally, it can also be observed that genres such as 'Pop' and 'Rock' feature a wide variety of colors in their cover images. These genres have more diverse subcategories and styles than other genres, resulting in such color variation. Note that due to text on cover images and background colors, grayscale colors are prominently displayed in the histogram. Therefore, Table 1 lists the representative colors most frequently used in cover images, excluding grayscale colors. As shown in Table 1, the most frequently used colors vary by music genre. Our proposed color palette estimation framework is designed based on this insight.

5. COLOR PALETTE ESTIMATION FRAMEWORK

Using the close relationship between musical pieces and colors used in their corresponding music cover images, we propose a framework designed to estimate appropriate color palettes for musical pieces. Fig. 4 shows an overview of our framework. We utilize an audio feature extractor and a color palette estimator to estimate a color palette for each musical piece. The audio feature extractor extracts audio features from musical pieces. To leverage recent advancements in audio models for downstream tasks, we use a pretrained audio model as the audio feature extractor. Then, the color palette estimator takes the extracted audio features as input and estimates appropriate color palettes for musical pieces.

To train the color palette estimator, we construct a large private dataset of musical pieces and their corresponding cover images, but we need the ground-truth color palettes to be estimated. Therefore, we extract the ground-truth color palette from each cover image by leveraging the data-driven color manifolds [14] described in Section 3 for our color palette extraction method.

5.1 Audio Feature Extractor

Audio models trained on large datasets have demonstrated their capabilities in downstream tasks [15, 23]. For example, the outputs of the final layer of an audio model are utilized in classification tasks, while audio embeddings are used in generative tasks. In our approach, we leverage the pretrained audio model $\mathcal{F}$ to extract audio features from musical pieces.

Let $\mathbf{A} = \{\mathbf{a}_n \in \mathbb{R}^T\}_{n=1}^N$ be a set of musical pieces, where T is the length of each musical piece and N is the number of musical pieces. Next, let $\mathbf{Z} = \{\mathbf{z}_n \in \mathbb{R}^d\}_{n=1}^N$ be a set of audio features, where d is the number of dimensions of each audio feature. The audio feature $\mathbf{z}_n$ can be extracted from the musical piece $\mathbf{a}_n$ by using the pretrained audio model $\mathcal{F}$ as follows:

$$\mathbf{z}_n = \mathcal{F}(\mathbf{a}_n). \tag{1}$$

We utilize the pretrained audio model with all of its trainable parameters fixed.

5.2 Color Palette Estimator

To estimate color palettes from the extracted audio features $\mathbf{z}_n$, we propose the color palette estimator $\mathcal{G}$, which consists of three linear layers with a GELU function [24].

Let $\mathbf{C} = \{\mathbf{c}_n \in \mathbb{R}^{3\times M}\}_{n=1}^N$ be a set of color palettes, where M is the number of colors in each color palette and each color is represented by a set of three numerical values (i.e., RGB values). The color palette $\mathbf{c}_n$ can be estimated from the audio feature $\mathbf{z}_n$ by using the color palette estimator $\mathcal{G}$ as follows:

$$\begin{aligned}\mathbf{c}_n &= \mathcal{G}(\mathbf{z}_n) \\ &= \sigma(W_3\mathrm{GELU}(W_2\mathrm{GELU}(W_1\mathbf{z}_n + \mathbf{b}_1) + \mathbf{b}_2) + \mathbf{b}_3),\end{aligned} \tag{2}$$

where σ is a sigmoid function. The parameters of the color palette estimator $\mathcal{G}$ are defined by $W_1 \in \mathbb{R}^{h\times d}, W_2 \in \mathbb{R}^{h\times h}, W_3 \in \mathbb{R}^{3\times M\times h}, \mathbf{b}_1 \in \mathbb{R}^h, \mathbf{b}_2 \in \mathbb{R}^h$, and $\mathbf{b}_3 \in \mathbb{R}^{3\times M}$, where the dimension of the hidden layer h is set to 768. While training, a dropout with a probability of 0.2 is applied to each output of the GELU functions.

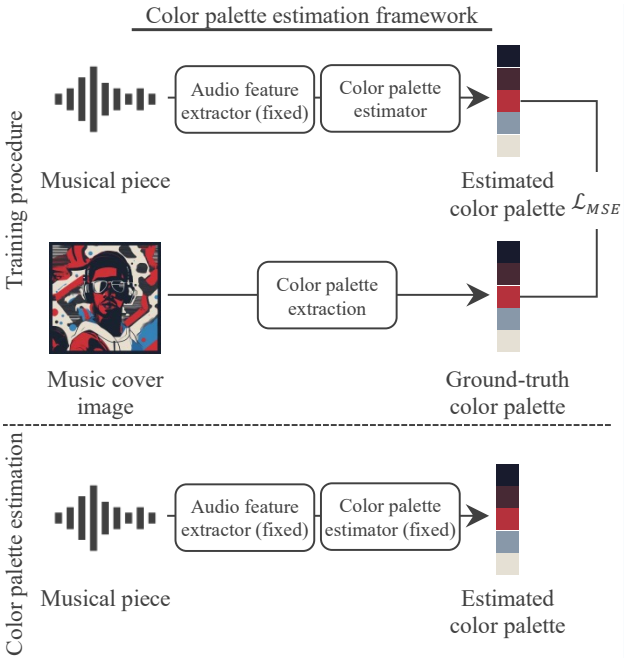

Figure 4. Overview of our proposed color palette estimation framework. **(Training procedure)** We start with an original pair of a musical piece and its corresponding music cover image. The musical piece is processed by a fixed audio model to extract its audio feature, and then a color palette estimator is trained to estimate a color palette from the audio feature. For this training, the ground-truth color palette is extracted from the cover image by using the color palette extraction method. We use the mean squared error (MSE) loss function to optimize our color palette estimator. **(Color palette estimation)** After training, the color palette estimator can be used to estimate appropriate color palettes for musical pieces.

5.3 Experimental Setup

5.3.1 Dataset

The large private dataset for training our color palette estimator contains music audio excerpts (each excerpt is a 30 s audio preview for trial listening, with a 44.1 kHz sampling rate) and their corresponding cover images (each image is an RGB image). The excerpts and their cover images are limited to single tracks, i.e., an original pair of an excerpt and its corresponding cover image is unique. The dataset contains 582,894 pairs of an excerpt and its corresponding cover image by 115,113 artists. We randomly split the dataset into training, validation, and test sets with an eight-one-one ratio (i.e., 466,316 pairs for the training set and 58,289 pairs for validation and test sets, respectively) and with no artists overlapping across these sets.

5.3.2 Implementation Details

As described in Section 4.1.2, we utilized the RGB color space for color representation, resized the cover images

into 224 px × 224 px, and normalized their RGB values to [0, 1]. We used a single NVIDIA A6000 GPU to train the color palette estimator. Our implementation was based on PyTorch [25]. We used the mini-batch size of 2,048. To train the color palette estimator, we used the Adam optimizer [26] with a learning rate of 1.0×10^{-4}. We calculated the mean squared error (MSE) loss function $\mathcal{L}_{MSE}$ between the estimated and ground-truth color palettes to optimize the parameters of the color palette estimator.

5.4 Experimental Settings

To clarify which audio feature extractor would be most effective in estimating the color palette, we conducted comparative experiments to investigate the importance of selection. Additionally, as a reference, we used color palettes composed of random colors.

5.4.1 Audio feature extractor

To extract audio features from musical pieces, we compared two audio models: *CLAP* [15] and *AudioToken* [16].

To use CLAP [15], we set the parameters of pretrained models available at HuggingFace's Transformers [27] (i.e., "laion/clap-htsat-fused"). Each musical piece was converted to a mel spectrogram through a CLAP feature extractor, and the CLAP audio model used the spectrogram as input. We fixed all parameters of the model. By using the model, we can obtain a 768-dimensional feature vector for each musical piece.

We also used AudioToken [16], which consists of a *bidirectional encoder representation from audio transformers (BEATs) [23]* model and an *embedder* [16] model. We set the parameters of pretrained models available at official GitHub repositories [2]. All parameters of the models were fixed. By employing the models, we can obtain a 768-dimensional feature vector for each musical piece.

5.4.2 Number of Colors in Color Palette

We used the number of colors $M = \{1, 2, 3, 4, 5\}$ in the color palettes for the experiments. To extract the color palettes from the cover images, we used the data-driven color manifolds [14] as described in Section 3.

5.5 Evaluation Metric for Comparative Experiments

In our comparative experiments, we used the *minimum color difference model (MCDM)* [28], which is practically designed to evaluate the color difference between two color palettes. The MCDM compares the two color palettes, each consisting of M colors, to determine their average color difference. First, the colors in the color palettes are converted from RGB to CIELAB [29]. Then, the MCDM calculates a CIELAB color difference between each color in one palette and all colors in the other palette, identifying the closest color match for each and recording the minimum differences. While multiple variants of CIELAB color difference exist, we here adopted the CIE1976 color difference [30] for evaluation. This process is repeated for every color in the first palette, resulting in M color difference values, which are then averaged to obtain a mean value, denoted as m_1. The same process is repeated for the second palette, finding the closest matches in the first palette and averaging the M minimum differences to obtain another mean value, m_2. Finally, the average of m_1 and m_2 gives the overall color difference between the two palettes. The lower the MCDM score, the closer the two color palettes. We leverage this MCDM to compare a color palette estimated with each experimental setting and a ground-truth color palette.

[2] The pretrained models are available at `https://github.com/microsoft/unilm/tree/master/beats` for the BEATs model and `https://github.com/guyyariv/AudioToken` for the embedder model.

Table 2. Results for the MCDM score on the test set of our dataset. A lower MCDM score indicates a closer match between the estimated color palettes and the ground-truth color palettes.

Audio feature extractor	M	MCDM score
CLAP	1	**28.37**
	2	**25.66**
	3	22.68
	4	21.72
	5	21.17
AudioToken	1	28.40
	2	25.69
	3	**22.66**
	4	**21.71**
	5	**21.14**
(Random)	1	69.29
	2	68.09
	3	66.35
	4	65.90
	5	65.50

5.6 Results

Table 2 presents the results for the MCDM score under each experimental setting. As shown in Table 2, our proposed framework achieves a much lower (i.e., better) MCDM score compared to random color palettes, with an improvement of over 40 points. This demonstrates the effectiveness of our framework. Additionally, these results support that there is a relationship between musical pieces and the color palettes used for designing their corresponding cover images because our framework succeeds in training the color palette estimator.

Regarding the selection of each experimental setting, there is no performance difference between the CLAP and AudioToken audio models, as shown in Table 2. This suggests that either audio model can be selected based on the intended application. As these results demonstrate, our framework effectively estimates appropriate color palettes for musical pieces.

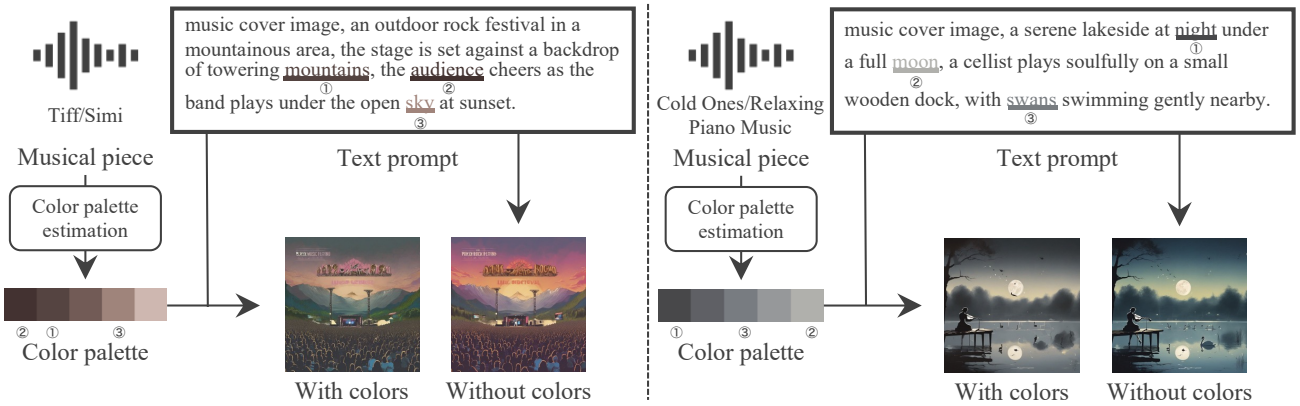

Figure 5. Example results of music cover image generation using the estimated color palettes. The upper left of each example shows the title and artist name of the input musical piece, the lower left shows the estimated color palette, the upper right shows the input text prompt, and the lower right shows the images generated using the input text prompt with/without the estimated color palette. In each text prompt, the colored words indicate objects being specified with that color by a user by leveraging the estimated color palette.

6. APPLICATION

Our framework can estimate color palettes appropriate for musical pieces, opening new doors in designing visual content for music content. In the context of generative AI, color palettes serve as an important means of controllability in image synthesis. Providing color hints allows users to align generative model's outputs with design intentions, thereby highlighting the importance of color palettes as compact and interpretable information in human–AI collaborative creativity [31, 32]. We showcase the potential of our framework through an example application for creating music cover images. This application leverages a cutting-edge *rich*-text-to-image model proposed by Ge et al. [33, 34], which enables precise color rendering on target objects or regions in image generation. By integrating our framework with this rich-text-to-image model, we can enhance the graphic design of cover images with color palettes. Instead of using text prompts such as "red", "blue", or "green", this application can render the generated images with specified color values, such as RGB values or Hex color codes. As described in Section 1, the advantage of being able to use specified color values is that color characteristics can be reflected in the cover image.

To use this application based on the rich-text-to-image model, a user needs to prepare a text prompt (i.e., a *plain text prompt*) that illustrates the intended cover images, in addition to the color palette that is estimated by using our color palette estimation framework. The user then specifies the objects' or regions' colors using the color palette, and its specified *rich text prompt* is given to the rich-text-to-image model in JSON format [3]. The rich-text-to-image model utilizes a general text-to-image model [4] to generate initial images, which are then refined through coloring. The color attributes used in the rich-text prompts control the colors of target objects or regions within the initial images, thereby improving the visual fidelity of the generated results.

The example results in Fig. 5 highlight the precision of color matching in cover images generated by the text-to-image model using text prompts and color palettes. A single text is displayed beneath the two generated images, as it is shared by both the plain text prompt and the rich text prompt. In the rich text prompt, the colored words indicate that those objects are specified with the corresponding colors, whereas in the plain text prompt, such color information is not provided. These examples demonstrate the potential of our color palette estimation framework, promising a wide array of applications, including music video creation and performance lighting design.

7. CONCLUSION

Our study finds the relationship between auditory and visual expression, specifically through the colors chosen for music cover images. Our findings suggest that there is a deliberate, meaningful selection of color palettes that reflect the essence of the music they express.

Our proposed framework, which utilizes the large private dataset encompassing a variety of genres, employs deep learning techniques to estimate appropriate color palettes for cover images. This bridges the fields of music computing and graphic design, especially for designers seeking to encapsulate the auditory experience of music in visual form. Our framework can streamline the design process by automating or helping the process of color selection.

Moreover, the example application of our framework, as demonstrated through the generation of cover images using the color palettes, underscores its effectiveness and potential impact on graphic design for music content. It offers a novel approach to designing cover images that are both aesthetically pleasing and deeply connected to a musical piece.

[3] Example rich text prompts are available at `https://github.com/SongweiGe/rich-text-to-image`. A plane text prompt can be converted into a rich text prompt in the JSON format by using the translator available at `https://rich-text-to-image.github.io/rich-text-to-json.html`.

[4] We used Stable Diffusion XL available at `https://huggingface.co/stabilityai/stable-diffusion-xl-base-1.0`

8. ACKNOWLEDGMENTS

This work was supported in part by JST CREST Grant Number JPMJCR20D4 and JSPS KAKENHI Grant Number 22K18017, Japan.

9. REFERENCES

[1] S. Oramas, F. Barbieri, O. Nieto, and X. Serra, "Multimodal deep learning for music genre classification," *Transactions of the International Society for Music Information Retrieval*, vol. 1, no. 1, pp. 4–22, 2018.

[2] J. Lībeks and D. Turnbull, "You can judge an artist by an album cover: Using images for music annotation," *IEEE MultiMedia*, vol. 18, no. 4, pp. 30–37, 2011.

[3] C. Ware, *Information visualization: perception for design*. Morgan Kaufmann, 2019.

[4] P. Valdez and A. Mehrabian, "Effects of color on emotions." *Journal of Experimental Psychology: General*, vol. 123, no. 4, pp. 394–409, 1994.

[5] L.-C. Ou, M. R. Luo, A. Woodcock, and A. Wright, "A study of colour emotion and colour preference. Part I: Colour emotions for single colours," *Color Research & Application*, vol. 29, no. 3, pp. 232–240, 2004.

[6] U. Ansorge and S. I. Becker, "Contingent capture in cueing: the role of color search templates and cue-target color relations," *Psychological Research*, vol. 78, pp. 209–221, 2014.

[7] M. Vad, "The album cover," *Journal of Popular Music Study*, vol. 33, no. 3, pp. 11–15, 2021.

[8] C. Seker, "New classics: The analysis of classical music album covers' digital age characteristics," *European Scientific Journal*, pp. 163–174, 2017.

[9] G. Friconnet, "A k-means clustering and histogram-based colorimetric analysis of metal album artworks: The colour palette of metal music," *Metal Music Studies*, vol. 9, no. 1, pp. 77–100, 2023.

[10] A. Dorochowicz and B. Kostek, "Relationship between album cover design and music genres," in *Proceedings of the 2019 Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA)*, 2019, pp. 93–98.

[11] J. Itten, *The elements of color*. John Wiley & Sons, 1970, vol. 4.

[12] T. L. Stone, S. Adams, and N. Morioka, *Color design workbook: A real world guide to using color in graphic design*. Rockport Pub, 2008.

[13] Y. Li and A. Sheopuri, "Creative design of color palettes for product packaging," in *Proceedings of the 2015 IEEE International Conference on Multimedia and Expo (ICME)*, 2015, pp. 1–6.

[14] C. H. Nguyen, T. Ritschel, and H.-P. Seidel, "Data-driven color manifolds," *ACM Transactions on Graphics (TOG)*, vol. 34, no. 2, pp. 1–9, 2015.

[15] Y. Wu, K. Chen, T. Zhang, Y. Hui, T. Berg-Kirkpatrick, and S. Dubnov, "Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation," in *Proceedings of the 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[16] G. Yariv, I. Gat, L. Wolf, Y. Adi, and I. Schwartz, "Adaptation of text-conditioned diffusion models for audio-to-image generation," in *Proceedings of the 24th Annual Conference of the International Speech Communication Association (Interspeech)*, 2023, pp. 5446–5450.

[17] A. W. Wells, "Music and visual color: A proposed correlation," *Leonardo*, vol. 13, pp. 101–107, 1980.

[18] J.-C. Wang, Y.-H. Yang, H.-M. Wang, and S.-K. Jeng, "The acoustic emotion gaussians model for emotion-based music annotation and retrieval," in *Proceedings of the 20th ACM International Conference on Multimedia (MM)*, 2012, pp. 89–98.

[19] J. De Berardinis, A. Cangelosi, and E. Coutinho, "The multiple voices of musical emotions: Source separation for improving music emotion recognition models and their interpretability," in *Proceedings of the 21st Conference of the International Society for Music Information Retrieval Conference (ISMIR)*, 2020, pp. 310–317.

[20] M. Pesek, P. Godec, M. Poredos, G. Strle, J. Guna, E. Stojmenova, M. Pogacnik, and M. Marolt, "Introducing a dataset of emotional and color responses to music." in *Proceedings of the 15th Conference of the International Society for Music Information Retrieval Conference (ISMIR)*, 2014, pp. 355–360.

[21] T. Kohonen, "The self-organizing map," *Proceedings of the IEEE*, vol. 78, no. 9, pp. 1464–1480, 1990.

[22] Y.-C. Hu and M.-G. Lee, "K-means-based color palette design scheme with the use of stable flags," *Journal of Electronic Imaging*, vol. 16, no. 3, p. 033003, 2007.

[23] S. Chen, Y. Wu, C. Wang, S. Liu, D. Tompkins, Z. Chen, and F. Wei, "Beats: Audio pre-training with acoustic tokenizers," in *Proceedings of the 40th International Conference on Machine Learning (ICML)*, 2023, pp. 5178–5193.

[24] D. Hendrycks and K. Gimpel, "Gaussian error linear units (gelus)," *arXiv preprint arXiv:1606.08415*, 2016.

[25] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "PyTorch: An imperative style, high-performance deep learning library," in *Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPS)*, vol. 32, 2019, pp. 8024–8035.

[26] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*, 2015, pp. 1–13.

[27] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, and A. M. Rush, "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations (EMNLP-SD)*, 2020, pp. 38–45.

[28] Q. Pan and S. Westland, "Comparative evaluation of color differences between color palettes," in *Proceedings of the 26th IS&T Color and Imaging Conference (CIC)*, vol. 26, 2018, pp. 110–115.

[29] T. Smith and J. Guild, "The cie colorimetric standards and their use," *Transactions of the Optical Society*, vol. 33, no. 3, p. 73, 1931.

[30] A. R. Robertson, "The CIE 1976 color-difference formulae," *Color Research & Application*, vol. 2, no. 1, pp. 7–11, 1977.

[31] V. Bozic, A. Djelouah, Y. Zhang, R. Timofte, M. Gross, and C. Schroers, "Versatile vision foundation model for image and video colorization," in *Proceedings of the ACM SIGGRAPH 2024 Conference Papers*, 2024, pp. 1–11.

[32] J. Yun, S. Lee, M. Park, and J. Choo, "iColoriT: Towards propagating local hints to the right region in interactive colorization by leveraging vision transformer," in *Proceedings of the 2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2023, pp. 1787–1796.

[33] S. Ge, T. Park, J.-Y. Zhu, and J.-B. Huang, "Expressive text-to-image generation with rich text," in *Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 7545–7556.

[34] ——, "Expressive text-to-image generation and editing with rich text," *International Journal of Computer Vision (IJCV)*, vol. 133, no. 7, pp. 4604–4622, 2025.