A COVER IMAGE COMPOSITION METHOD FOR MUSIC CONTENT USING MULTIMODAL IMAGE RETRIEVAL AND CROPPING TECHNIQUES

Takayuki NAKATSUKA, **Masahiro HAMASAKI**, and **Masataka GOTO**

National Institute of Advanced Industrial Science and Technology (AIST), Tsukuba, Japan
{takatuki.nakatsuka, masahiro.hamasaki, m.goto}@aist.go.jp

ABSTRACT

Aesthetic image compositions capture people's attention. For music content, music cover images and thumbnail images play an important role in enhancing visibility and viewer engagement of the content. While image cropping techniques can generate aesthetic image compositions from a single image, generating image compositions from multiple images for music content remains unexplored. The challenge is to ensure that these image compositions maintain a specified aspect ratio, possess high aesthetic quality, and are contextually relevant to the music content. To address this challenge, we propose a cover image composition method that generates aesthetic image compositions from an image collection and retrieves image compositions suitable for music content. The key technical aspect involves constructing a multimodal embedding space using multimodal image retrieval and cropping techniques. Within this space, feature vectors of music audio, aesthetic image compositions, and text with similar contexts can be placed closely by optimizing them using our proposed loss function. Given music content (music audio or text) as a query, our method can retrieve image compositions suitable for the query on the basis of the similarities of their feature vectors in the space. Through qualitative analysis and quantitative evaluation, we demonstrate the effectiveness of our proposed method.

1. INTRODUCTION

Visual content designed for music, such as cover and thumbnail images, serves as the initial point of engagement between music content and viewers, substantially contributing to advertising music content and enriching the viewing experience [1, 2]. Consequently, musicians devote their efforts to creating aesthetic cover (or thumbnail) images tailored to their music content. To create such images, some leverage stock images from sources like Adobe Stock [3] and Unsplash [4]. The problem is that these stock images do not always have the optimal aspect ratio required for such images, typically specified as 1:1 for cover images and 16:9 for thumbnail images by online music services. To address this problem, image cropping techniques are used to generate aesthetic image compositions in the specified aspect ratio.

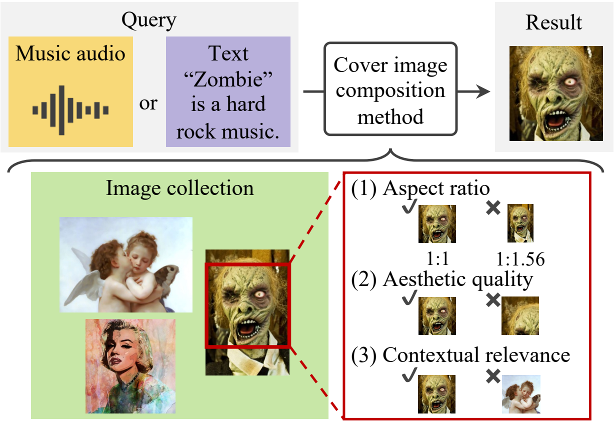

Figure 1. Conceptual design of the task in this study. Our objective is to retrieve image compositions suitable for a music audio or text query. The image compositions are generated from an image collection and fulfill three requirements: they have a specified aspect ratio, high aesthetic quality, and contextual relevance to the query.

Image cropping can identify aesthetic image compositions within an image, considering factors such as saliency, composition, and aesthetics [5]. Several studies have proposed image cropping techniques aimed at selecting specific regions or objects within an image to align with the user's intention [6–9]. However, these techniques primarily focus on cropping from a single image and do not assist users in finding images from an image collection, such as stock images. Given such a collection, generating aesthetic image compositions suitable for music content remains challenging.

In this paper, we propose a cover image composition method to address this challenge, as illustrated in Figure 1. Our method leverages music content (music audio or text) as a query to retrieve suitable image compositions. Such a query enables users to easily find their desired image compositions. Furthermore, cover and thumbnail images must be adjusted to a specified aspect ratio, possess a high aesthetic quality, and match the context of the music content. Therefore, our method takes these requirements into account when generating image compositions.

To generate such image compositions, our method utilizes multimodal image retrieval and cropping techniques. For the retrieval, we first construct a multimodal embedding space in which each music audio, image, and text is represented as a feature vector by training respective encoders that embed them into the space.

To train the encoders, our method uses contrastive learning [10–12], a deep learning technique that embeds data into the multimodal embedding space so that feature vectors of similar context data are close to each other, while those of dissimilar ones are far away. By optimizing the space using contrastive learning, we can closely align the feature vectors of music audio, images, and text with similar contexts, thereby facilitating the search for image compositions on the basis of a music audio (or text) query.

Additionally, we use image outpainting and cropping techniques to ensure that the encoders identify images of high aesthetic quality. Image outpainting can seamlessly expand the boundary of an input image. We utilize the mask-aware transformer (MAT) [13] to generate the outpainted images of cover (or thumbnail) images. We then train the encoders to prioritize cover (or thumbnail) images over images randomly cropped from the outpainted images in the multimodal embedding space.

The trained encoders can be utilized to embed a music audio (or text) query and image compositions into the multimodal embedding space. We use a grid-anchor formulation [14] to generate image composition candidates from each image in the image collection. Subsequently, we can generate a ranked list of image composition candidates on the basis of the similarity between the feature vectors of the query and each candidate.

We qualitatively show the effectiveness of our method using the TAD66K dataset [15] as the image collection. In addition, we demonstrate our method's effectiveness through quantitative evaluation on the public YT8M-MusicVideo dataset [16] and the private Album Songs 5 Million (AS5M) dataset.

2. RELATED WORK

Image cropping automatically selects visually-appealing regions or objects from an image for various applications [17]. Advancements in deep learning techniques have enabled image cropping techniques to become more practical [18]. However, these techniques mainly focus on improving the aesthetic quality of the cropped images, resulting in the images that do not capture the user's intention.

Several studies have been pursuing the potential for further advancement in image cropping, particularly in capturing the user's intention through multimodal input [6–9]. Santella et al. introduced a framework that implicitly utilizes gaze-based interactions to identify accurate regions of interest [6]. Bhattacharya et al. developed a framework that recomposes an image on the basis of the user-selected foreground object [7]. Horanyi et al. proposed caption and aesthetic-guided image cropping, which leverages pretrained models for image captioning and aesthetic tasks [8]. Zhong et al. developed a framework that integrates OWL-ViT [19] and DETR [19] to achieve query-conditioned image cropping [9]. The drawback of these image cropping techniques is that they only search for image compositions in a single image. Our method can generate image compositions from multiple images, facilitating the search for image compositions the users want.

3. PROPOSED METHOD

This section describes our proposed method. Figure 2 shows an overview of our proposed method.

3.1 Data Representation

Our method deals with three modalities of data: music audio, images, and text. Here, we describe each data representation in our method.

3.1.1 Music Audio Representation

The music audio is converted to a mel spectrogram through a feature extractor of contrastive language-audio pretraining (CLAP) [20] available in Transformers [21], and our audio encoder is trained using the spectrogram as input. To train our audio encoder, we apply a masking technique [22] and a random crop technique [23] to the spectrogram for data augmentation. The masking technique generates random masks on the spectrogram in both frequency and time domains, and the random crop technique selects a random section of the music audio.

3.1.2 Image Representation

We use an RGB image resized to $224\,\mathrm{px} \times 224\,\mathrm{px}$ as input for our image encoder. Since most datasets for image cropping are single-modal, comprising only images, multimodal datasets that include music audio are not readily available. Hence, we create such a dataset using image outpainting and cropping techniques. Specifically, we utilize MAT [13], an outpainting technique, to expand the boundary of the original cover (or thumbnail) image. We resize the image such that its longer side becomes 224 pixels while maintaining the original aspect ratio, and then we center the resized image and use MAT to outpaint around it, resulting in an image size of $512\,\mathrm{px} \times 512\,\mathrm{px}$. During training, we operate under the assumption that the region of the original image represents a more suitable composition than randomly cropped regions from the outpainted image. This approach is based on the actual design workflow in which cover and thumbnail images are often created by cropping from photos or illustrations. It has the unprecedented advantage that original images actually cropped by professionals can be used as correct answers in this manner. To train our image encoder, we use both the original and outpainted images. The outpainted images undergo a series of data augmentation: a random resized crop (with scale range [0.57, 1.43] and ratio range [0.4, 2.5]), a random horizontal flip (with a probability of 0.5), and random erasing (with a probability of 0.2) [24].

3.1.3 Text Representation

We tokenized text generated by using a keyword-to-caption augmentation technique [20] with a maximum length of 77, which is the same setup as contrastive language-image pretraining (CLIP) [25]. To train our

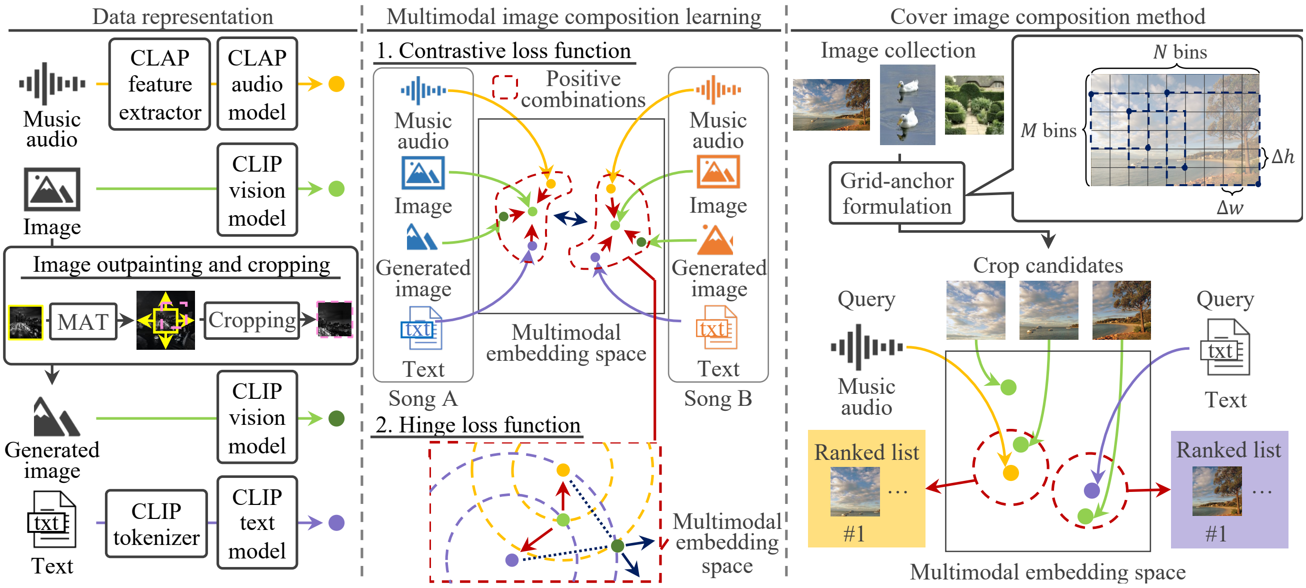

Figure 2. Overview of the proposed method that leverages multimodal image retrieval and cropping techniques. Our method utilizes music audio, images, and text. In multimodal image composition learning, we train encoders so that feature vectors of music audio, an image, and text from the same song are close in a multimodal embedding space, whereas those from different songs are far apart. Additionally, we use outpainting and cropping techniques to generate images. We also train the encoders so that the feature vector of an aesthetic (i.e., original) image is closer to those of the corresponding music audio and text than that of the generated image in the space. Once the training is completed, our cover image composition method can generate a ranked list of image composition candidates for a given query from an image collection.

text encoder, keywords corresponding to metadata are randomly dropped [26] at a ratio of 0.05 for each metadata.

3.2 Multimodal Image Composition Learning

This section describes our designs of each encoder and a learning framework to train the encoders.

3.2.1 *Encoder Architecture*

We use an audio model of CLAP [20] as the audio encoder, and use image and text models of CLIP [25] as the image and text encoders, respectively. We utilized pretrained models available in Transformers [21] (i.e., "laion/clap-htsat-fused" for CLAP (audio model) and "openai/vit_base_patch16_224" for CLIP (image and text models)). During training, we updated the projection layers of the encoders. Each encoder outputs a 512-dimensional feature vector.

3.2.2 *Learning Framework*

To jointly train multiple modalities of data, contrastive learning is an effective approach [16, 20, 25]. Inspired by contrastive loss functions that calculate the loss on the basis of N-pairs of instances (i.e., one positive pair and $N-1$ negative (or irrelevant) pairs) like N-pairs loss [10], InfoNCE loss [11], and MoCo [12], we design a contrastive loss function $\mathcal{L}_C$ that considers all pairwise combinations of music audio, images, and text as follows:

$$\mathcal{L}_C = -\frac{1}{m}\frac{1}{|\mathcal{S}|}\sum_{\boldsymbol{S}\in\mathcal{S}}\sum_{i=1}^{m}\log\frac{e^{S_{i+}/\tau}}{\sum_{j=1}^{m}e^{S_{ij}/\tau}}, \qquad (1)$$

where m is a mini-batch size, τ is a temperature scaling parameter that controls the scale of the loss function, $+$ indicates a positive instance of an anchor, $\boldsymbol{S}$ is a similarity matrix whose element S_{ij} is defined as the cosine similarity $\mathrm{sim}(\cdot,\cdot)$ between an anchor and an instance in a mini-batch, and $\mathcal{S}$ is a set of similarity matrices that are constructed by using all combinations of music audio feature vectors $\{\boldsymbol{z}_n^A\}_{n=1}^m$, image feature vectors $\{\boldsymbol{z}_n^I\}_{n=1}^m$, and text feature vectors $\{\boldsymbol{z}_n^T\}_{n=1}^m$. While training, we also utilize feature vectors of randomly cropped images, $\{\boldsymbol{z}_n'^I\}_{n=1}^m$, as described in Section 3.1. The contrastive loss function is effective for embedding contextually irrelevant data that are far apart in the multimodal embedding space.

Furthermore, to facilitate the training of the encoders to learn aesthetic image compositions, we design a hinge loss function $\mathcal{L}_H$ inspired by the margin ranking loss [27, 28] as follows:

$$\begin{aligned}\mathcal{L}_H = &\sum_{i=1}^{m}\max\{0, \alpha - \mathrm{sim}(\boldsymbol{z}_i^A, \boldsymbol{z}_i^I) + \mathrm{sim}(\boldsymbol{z}_i^A, \boldsymbol{z}_i'^I)\}\\ &+\sum_{i=1}^{m}\max\{0, \alpha - \mathrm{sim}(\boldsymbol{z}_i^T, \boldsymbol{z}_i^I) + \mathrm{sim}(\boldsymbol{z}_i^T, \boldsymbol{z}_i'^I)\},\end{aligned} \qquad (2)$$

where α is a margin. The hinge loss function is effective for embedding the original image closer to the feature vector of the corresponding music audio (or text) than that of the randomly cropped images in the space.

To leverage the advantages of both the contrastive and hinge loss functions, we formulate a novel loss function that incorporates those loss functions as follows:

$$\mathcal{L} = \mathcal{L}_C + \lambda_H\mathcal{L}_H, \qquad (3)$$

where λ_H is a weight.

3.3 Cover Image Composition Method

We use the grid-anchor formulation [14] to generate image composition candidates with a specified aspect ratio from each image in the image collection. First, image composition candidates are automatically determined by using the grid-anchor formulation that constructs an image grid with $M \times N$ bins on the original image. Each bin has a width of Δw px and a height of Δh px. Then, the grid-anchor formulation chooses a region to be cropped by selecting two anchor points so that the region is at least 40 percent of the size of the original image. This formulation can efficiently determine image composition candidates of a specified aspect ratio from the original image. We use $\Delta w = \Delta h = 12$ when cropping an image with an aspect ratio of 1:1 and $\Delta w = 16, \Delta h = 9$ for 16:9. The values of M and N are automatically determined from the size of each bin and the aspect ratio of the original image. As a result, we can obtain a dozen to several hundred image composition candidates of a specified aspect ratio from each image in the image collection. Note that the number of image composition candidates generated by the proposed method depends on the size of each image.

Then, we can calculate how an image composition candidate matches a query as follows:

$$\mathcal{F}_i = \mathrm{sim}(\boldsymbol{z}^Q, \boldsymbol{z}_i^I) \tag{4}$$

where $\mathbf{z}^{\mathrm{Q}}$ and $\mathbf{z}_i^{\mathrm{I}}$ are the feature vectors of the query and the i-th instance of the image composition candidates, respectively. We calculate feature vectors of the query and image composition candidates using the trained encoders. By sorting the values $\{\mathcal{F}_i\}$ in descending order, we can obtain a ranked list of the image composition candidates.

4. EXPERIMENTS AND RESULTS

This section describes a qualitative analysis and comparison experiments to evaluate our method's effectiveness.

4.1 Dataset

We utilized three datasets (one image dataset and two multimodal datasets) in our experiments.

4.1.1 Image dataset

The **TAD66K dataset** [15] is specifically designed for image aesthetics assessment. The dataset contains over 66K images of various aspect ratios, covering 47 popular themes. We utilized this dataset as the image collection.

4.1.2 Multimodal dataset

The **YT8M-MusicVideo dataset** [16] is a subset of the YouTube-8M dataset [30], comprising videos tagged as "music video." We collected 73,113 triplets consisting of music audio (average length of 4 min with a 48 kHz sampling rate), the corresponding thumbnail image (an RGB image with an aspect ratio of 16:9), and the corresponding metadata including title, channel name, and upload date on YouTube from 60,785 YouTube channels. We randomly split the dataset into training (64,001 songs), validation (7,112 songs), and test (2,000 songs) sets with no channels overlapping across these sets.

The **AS5M dataset** is a private dataset containing triplets of a music audio excerpt (a 30 s audio preview for trial listening, with a 44.1 kHz sampling rate), the corresponding cover image (a square RGB image), and the corresponding metadata including song title, artist name, collection name, music genre, and release date. The dataset contains 5,920,828 music audio excerpts and their metadata by 174,629 artists, and 1,115,668 cover images. Because multiple excerpts from an album are associated with a single cover image, each image corresponds to about 5.3 excerpts on average. The excerpts, typically representative music sections, were already cropped on a music streaming service from which they were crawled. The corresponding cover images and metadata were crawled simultaneously. The songs encompass a variety of music genres (over 250, according to the streaming service). We randomly split the dataset into training, validation, and test sets with an eight-one-one ratio and with no artists or images overlapping across these sets. For evaluation, we constructed ten folds of test subsets by randomly selecting 2,000 triplets of a music audio excerpt, cover image, and text prompt for each fold from the test set.

We determined the size of each test set by following the setup used in related works [16, 31].

4.2 Qualitative Analysis

We performed a qualitative analysis to demonstrate that the proposed method can generate aesthetic image composition candidates and retrieve appropriate image compositions in response to each query. The most effective way to demonstrate the proposed method is to test it on real examples [16]. We thus used music audio available on YouTube [1] as music audio queries, and text available on Wikipedia [2] as text queries. The queries were selected on the basis of their high popularity, that is, hit charts (rankings) of music sales and YouTube views. For the image collection, we used all images in the TAD66K dataset [15].

Figure 3 shows example results of our method. The images in the top and bottom four rows are formatted in an aspect ratio of 16:9 and 1:1, respectively. From the ranked list, we list the top two images as "most matched", the middle three images as "moderately matched", and the bottom two as "least matched." The results show that our method captures the property of each query and retrieves image compositions that match either the music audio or text queries. For example, in the case of "White Christmas" (third row of Figure 3), the word "Christmas" resulted in the most matched images, such as the board shaped like Santa Claus, decorated Christmas tree, and star-shaped ornament. In the case of "Call Me Maybe" (fifth row of Figure 3), the top ranked results feature charming images since the song is categorized as teen-pop. Altogether,

[1] Each music audio query in Figure 3 can be accessed at https://youtu.be/{ID}.

[2] We used the first paragraph that introduces a song. Each text query in Figure 3 can be accessed at https://en.wikipedia.org/wiki/{Url}.

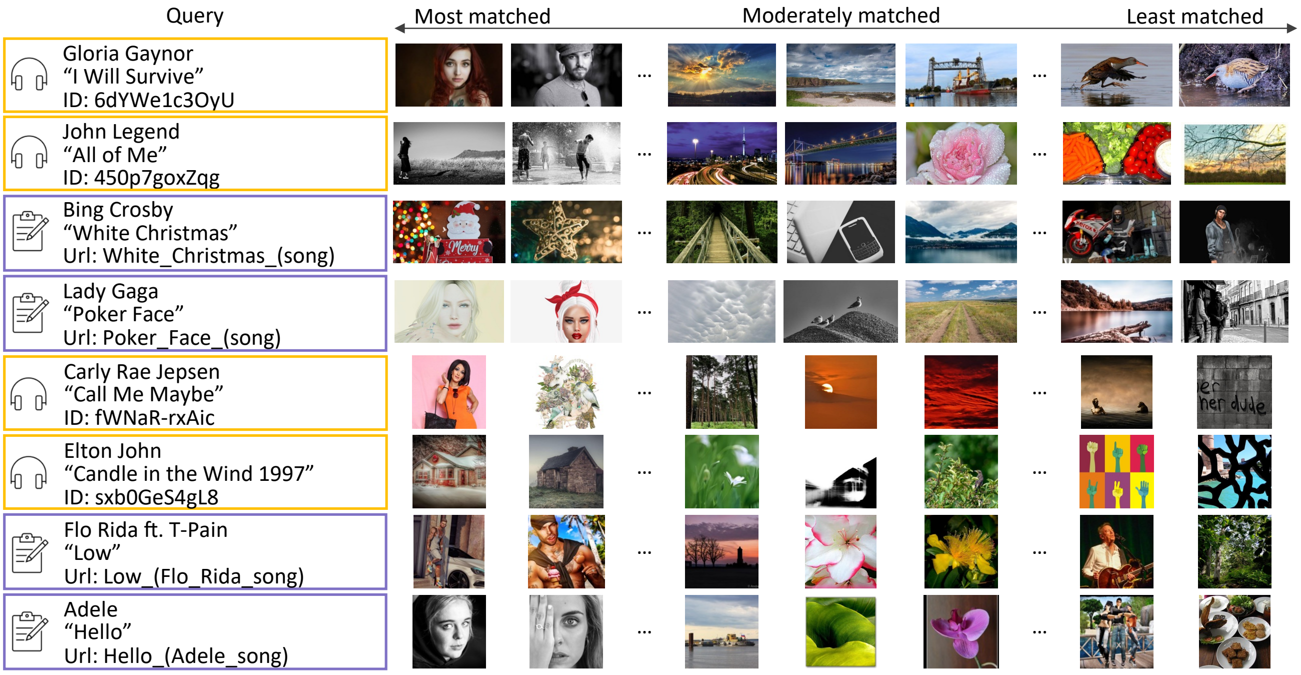

Figure 3. Example results, showcasing ranked lists of image composition candidates generated from the TAD66K dataset [29]. The query with the headphone icon 🎧 represents a music audio query (`https://youtu.be/6dYWe1c3OyU`, etc.), while the query with the text icon 📝 represents a text query (`https://en.wikipedia.org/wiki/White_Christmas_(song)`, etc.).

"most matched" images are similar in context, showing that the multimodal embedding space is well constructed. These results demonstrate that the proposed method can retrieve appropriate image compositions in response to each query.

4.3 Study on Multimodal Image Composition Learning

While training the encoders, we assumed that a combination of music audio, an image, and text from the same song is positive, whereas a combination from different songs is negative. In addition, we designed our loss function so that the aesthetic (i.e., original) image is closer to the corresponding music audio (or text) than the randomly cropped images in the multimodal embedding space. We thus quantitatively evaluated how closely the feature vectors of an original pair are located in the multimodal embedding space.

4.3.1 Experimental Setup

Training Details. Our implementation was based on PyTorch [32]. We used 16 NVIDIA V100 GPUs under each experimental condition. Each GPU computed 128 triplets of music audio, images, and text per iteration. When training the encoders, we used the Adam optimizer [33] with an initial learning rate of 1.0e-4. We set the weight λ_H to 1.0 and the margin α to 0.0. Following the setup in MoCo [11], we also set the temperature-scaling value τ to 0.07.

Evaluation Metrics. We used the recall@k (R@k), which is the standard evaluation metric in retrieval tasks, to assess the accuracy with which each method could find the original image corresponding to a music audio (or text) query from the image collection. R@k evaluates how much correct content is retrieved in the top results. A higher recall at a given k means that the retrieval method is more practical. We displayed R@k as a percentage.

In addition, we used the intersection over union (IoU), which is the standard evaluation metric in image cropping, to assess how precisely each method could identify the region of the original image within the outpainted image. For this evaluation, we used the original image and a hundred images randomly cropped from the outpainted images as the image composition candidates. We calculated the IoU between the region of the original image and that of the top ranked image composition retrieved by each method. We then averaged the obtained IoU scores over the test set.

4.3.2 Conditions

To demonstrate the effectiveness of our method, we compared the following four methods.

- **Baseline** solely used the contrastive loss function, where the similarity matrices only consist of cross-modal feature vectors (e.g., $\mathrm{sim}(\boldsymbol{z}_i^A, \boldsymbol{z}_j^I)$).
- **Baseline + Self-modal supervision** solely used the contrastive loss function, where the similarity matrices consist of self-modal feature vectors (e.g., $\mathrm{sim}(\boldsymbol{z}_i^A, \boldsymbol{z}_j^A)$) in addition to cross-modal feature vectors.
- **Baseline + Self-modal supervision + Outpainted images** solely used the contrastive loss function, where the similarity matrices consist of the feature vectors of randomly cropped images (e.g.,

Table 1. Results for R@k and IoU on the test set of the YT8M-MusicVideo dataset [16] with k set to 1, 5, and 10.

Method	Music-to-Image				Text-to-Image			
	R@1 ↑	R@5 ↑	R@10 ↑	IoU ↑	R@1 ↑	R@5 ↑	R@10 ↑	IoU ↑
Baseline	0.75	3.25	5.45	0.3415	3.45	8.10	10.60	0.4787
+ Self-modal supervision	0.75	3.35	5.95	0.3669	**3.70**	**8.15**	**10.75**	0.4872
+ Outpainted images	**0.85**	4.10	6.80	0.3722	3.60	7.95	10.60	0.5096
+ Hinge loss function (ours)	**0.85**	**4.30**	**7.15**	**0.4415**	3.55	8.00	**10.75**	**0.5271**

Table 2. Results for R@k and IoU on the test subsets of the AS5M dataset with k set to 1.

Method	Music-to-Image		Text-to-Image	
	R@1 ↑	IoU ↑	R@1 ↑	IoU ↑
Baseline	1.77 ± 0.27	0.4294 ± 0.009	6.57 ± 0.33	0.6679 ± 0.005
+ Self-modal supervision	2.22 ± 0.18	0.4524 ± 0.006	6.40 ± 0.34	0.6754 ± 0.006
+ Outpainted images	2.25 ± 0.26	0.4391 ± 0.005	6.39 ± 0.36	$\mathbf{0.6880 \pm 0.007}$
+ Hinge loss function (ours)	$\mathbf{2.44 \pm 0.25}$	$\mathbf{0.4766 \pm 0.008}$	$\mathbf{6.70 \pm 0.31}$	0.6817 ± 0.007

$\mathrm{sim}(\boldsymbol{z}_i^I, \boldsymbol{z}_j'^I)$) in addition to self- and cross-modal feature vectors.

- **Baseline + Self-modal supervision + Outpainted images + Hinge loss function**, which is our proposed method, used both the contrastive and hinge loss functions as described in Section 3.2.

To quantitatively evaluate the performance of each method, we set up two tasks, Music-to-Image and Text-to-Image, in which music audio and text were used as a query, respectively, to retrieve a corresponding image and identify a corresponding region.

4.3.3 Results

Table 1 shows that our method generally outperformed the baseline method by 0.1 points for R@1, 1.05 points for R@5, 1.7 points for R@10, and 0.1 points for IoU in the Music-to-Image task, and by 0.1 points for R@1, 0.15 points for R@10, and 0.0484 points for IoU in the Text-to-Image task on the YT8M-MusicVideo dataset. Likewise, Table 2 shows that our method outperformed the baseline method by 0.67 points for R@1 and 0.0472 points for IoU in the Music-to-Image task, and by 0.13 points for R@1 and 0.0138 points for IoU in the Text-to-Image task on the AS5M dataset.

These results demonstrate the effectiveness of our multimodal image composition learning. That is, our method can closely align the feature vector of the given query with that of the corresponding image of high aesthetic quality in the multimodal embedding space. Thus, our proposed method can effectively retrieve image compositions that are suitable for the query.

4.3.4 Visualization of Feature Vectors

We further investigated the nature of obtained feature vectors to clarify which query properties facilitate our proposed multimodal image composition learning. For example, Libeks et al. found that cover images were characterized by music genre [35]. Here, we explored music genre and artist tags. We used the test subsets of the AS5M dataset. The feature vectors were visualized by using t-SNE [34] to project them to a two-dimensional space.

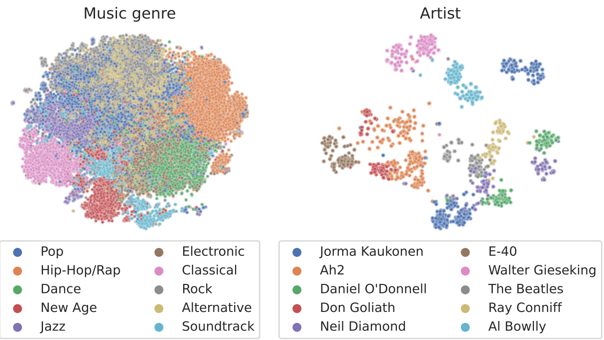

Figure 4. Visualization of music audio, image, and text feature vectors in the test subsets of the AS5M dataset with respect to music genre tags (left) and artist tags (right). These are visualized using t-SNE [34].

The left side of Figure 4 visualizes feature vectors by genre tags. We used the feature vectors of the music audio, images, and text for 12,180 songs in the top 10 most popular genres in the dataset. The result shows that feature vectors form clusters for each of several genres, such as Hip-Hop/Rap and Classical. That is, music audio, images, and text in each of these genres are closely associated with each other. The right side of Figure 4 visualizes feature vectors by artist tags. We used the feature vectors for 757 songs by the top 10 most popular artists in the dataset. Surprisingly, the result shows that feature vectors are clearly separated by each artist tag, while we simply assumed that a combination of music audio, an image, and text from the same song is positive. That is, music embodies the unique personality of each artist.

These results demonstrate that our proposed method successfully trains the encoders so that the feature vectors of music audio, images, and text with similar contexts are close to each other in the multimodal embedding space.

5. DISCUSSION

The proposed method is useful in various situations, e.g., when musicians want to add a cover (or thumbnail) image to their own music, using a large collection of images or photographs that they personally own or that are free of rights issues. The retrieved image composition can be used as is or further adjusted as a cover (or thumbnail) image.

In such a situation, one might wonder why not use recent multimodal generative models capable of generating cover images from various inputs [36]. A primary concern with generative techniques is the potential to replicate existing images from training data [37, 38]. For musicians, legal ambiguities and risks of using generated images can be a substantial drawback.

In contrast, the use of self-collected or owned images or commercial stock images is free of rights concerns and is likely to be the preferred option, especially for commercial use. Therefore, this paper focuses on the unique challenge of retrieving cover image compositions from an image collection without daring to consider using the generative models.

6. CONCLUSION

We proposed a method of retrieving image compositions suitable for a music audio (or text) query. The contributions of this paper can be summarized as follows. First, to ensure that image compositions have the specified aspect ratio, high aesthetic quality, and contextual relevance to the query, we utilized multimodal image retrieval and cropping techniques, which offer a novel solution to this challenge. Second, we demonstrated that the proposed method can retrieve image compositions suitable for queries, as our qualitative analysis shows. Third, our proposed method succeeded in training the encoders so that the feature vectors of music audio, images, and text from the same song (i.e., those with similar contexts) are close to each other in the multimodal embedding space, as demonstrated by our quantitative evaluation.

This work will lead to further development of assistive tools that can deal with various music content.

Acknowledgments

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

7. REFERENCES

[1] M. Vad, "The album cover," *J. Pop. Music Stud.*, vol. 33, no. 3, pp. 11–15, 2021.

[2] S. J. Cunningham and D. M. Nichols, "How people find videos," in *Proceedings of the ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL)*, 2008, pp. 201–210.

[3] A. Inc., "Adobe stock," 2023, retrieved March 1, 2024 from https://stock.adobe.com.

[4] I. Getty Images Holdings, "Unsplash," 2023, retrieved March 1, 2024 from https://unsplash.com.

[5] Y. Deng, C. C. Loy, and X. Tang, "Image aesthetic assessment: An experimental survey," *IEEE Signal Process. Mag.*, vol. 34, no. 4, pp. 80–106, 2017.

[6] A. Santella, M. Agrawala, D. DeCarlo, D. Salesin, and M. Cohen, "Gaze-based interaction for semi-automatic photo cropping," in *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)*, 2006, pp. 771–780.

[7] S. Bhattacharya, R. Sukthankar, and M. Shah, "A framework for photo-quality assessment and enhancement based on visual aesthetics," in *Proceedings of the ACM International Conference on Multimedia (MM)*, 2010, pp. 271–280.

[8] N. Horanyi, K. Xia, K. M. Yi, A. K. Bojja, A. Leonardis, and H. J. Chang, "Repurposing existing deep networks for caption and aesthetic-guided image cropping," *Pattern Recognit.*, vol. 126, p. 108485, 2022.

[9] Z. Zhong, M. Cheng, Z. Wu, Y. Yuan, Y. Zheng, J. Li, H. Hu, S. Lin, Y. Sato, and I. Sato, "ClipCrop: Conditioned cropping driven by vision-language model," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 294–304.

[10] K. Sohn, "Improved deep metric learning with multi-class n-pair loss objective," in *Proceedings of the Advances in Neural Information Processing Systems (NIPS)*, 2016, pp. 1857–1865.

[11] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[12] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 9729–9738.

[13] W. Li, Z. Lin, K. Zhou, L. Qi, Y. Wang, and J. Jia, "MAT: Mask-aware transformer for large hole image inpainting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 10 758–10 768.

[14] H. Zeng, L. Li, Z. Cao, and L. Zhang, "Grid anchor based image cropping: A new benchmark and an efficient model," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 3, pp. 1304–1319, 2022.

[15] G. Jia, H. Huang, C. Fu, and R. He, "Rethinking image cropping: Exploring diverse compositions from global views," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 2446–2455.

[16] D. Surís, C. Vondrick, B. Russell, and J. Salamon, “It's time for artistic correspondence in music and video,” in *Proceedings of the IEEE/CVF Computer Vision and Pattern Recognition Conference (CVPR)*, 2022, pp. 10 564–10 574.

[17] B. Suh, H. Ling, B. B. Bederson, and D. W. Jacobs, “Automatic thumbnail cropping and its effectiveness,” in *Proceedings of the ACM Symposium on User Interface Software and Technology (UIST)*, 2003, pp. 95–104.

[18] J. Zhang, Y. Miao, and J. Yu, “A comprehensive survey on computational aesthetic evaluation of visual art images: Metrics and challenges,” *IEEE Access*, vol. 9, pp. 77 164–77 187, 2021.

[19] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, “End-to-end object detection with transformers,” in *European Conference on Computer Vision (ECCV)*, 2020, pp. 213–229.

[20] 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 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[21] 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 Empirical Methods in Natural Language Processing: System Demonstrations (EMNLP-SD)*, 2020, pp. 38–45.

[22] Y. Luo and N. Mesgarani, “Conv-TasNet: Surpassing ideal time–frequency magnitude masking for speech separation,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 8, pp. 1256–1266, 2019.

[23] R. Takahashi, T. Matsubara, and K. Uehara, “Data augmentation using random image cropping and patching for deep CNNs,” *IEEE Trans. Circuits. Syst. Video Technol.*, vol. 30, no. 9, pp. 2917–2931, 2019.

[24] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, “Random erasing data augmentation,” in *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2020, pp. 13 001–13 008.

[25] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, “Learning transferable visual models from natural language supervision,” in *Proceedings of the International Conference on Machine Learning (ICML)*, 2021, pp. 8748–8763.

[26] T. Sellam, D. Das, and A. P. Parikh, “BLEURT: Learning robust metrics for text generation,” in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020, pp. 7881–7892.

[27] F. Schroff, D. Kalenichenko, and J. Philbin, “FaceNet: A unified embedding for face recognition and clustering,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 815–823.

[28] J. Wang, Y. Song, T. Leung, C. Rosenberg, J. Wang, J. Philbin, B. Chen, and Y. Wu, “Learning fine-grained image similarity with deep ranking,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 1386–1393.

[29] S. He, Y. Zhang, R. Xie, D. Jiang, and A. Ming, “Rethinking image aesthetics assessment: Models, datasets and benchmarks,” in *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 2022, pp. 942–948.

[30] S. Abu-El-Haija, N. Kothari, J. Lee, P. Natsev, G. Toderici, B. Varadarajan, and S. Vijayanarasimhan, “YouTube-8M: A large-scale video classification benchmark,” *arXiv preprint arXiv:1609.08675*, 2016.

[31] L. Prétet, G. Richard, and G. Peeters, “Cross-modal music-video recommendation: A study of design choices,” in *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*, 2021, pp. 1–9.

[32] 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 Conference on Neural Information Processing Systems (NeurIPS)*, 2019, pp. 8024–8035.

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

[34] L. Van der Maaten and G. Hinton, “Visualizing data using t-SNE,” *J. Mach. Learn. Res.*, vol. 9, no. 11, pp. 2579–2605, 2008.

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

[36] S. Wu, H. Fei, L. Qu, W. Ji, and T.-S. Chua, “NExT-GPT: Any-to-any multimodal LLM,” *arXiv preprint arXiv:2309.05519*, 2023.

[37] G. Somepalli, V. Singla, M. Goldblum, J. Geiping, and T. Goldstein, “Diffusion art or digital forgery? investigating data replication in diffusion models,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 6048–6058.

[38] ——, “Understanding and mitigating copying in diffusion models,” *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 36, pp. 47 783–47 803, 2023.