Content-Based Music-Image Retrieval Using Self- and Cross-Modal Feature Embedding Memory

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

Abstract

This paper describes a method based on deep metric learning for content-based cross-modal retrieval of a piece of music and its representative image (i.e., a music audio signal and its cover art image). We train music and image encoders so that the embeddings of a positive music-image pair lie close to each other, while those of a random pair lie far from each other, in a shared embedding space. Furthermore, we propose a mechanism called self- and cross-modal feature embedding memory, which stores both the music and image embeddings of any previous iterations in memory and enables the encoders to mine informative pairs for training. To perform such training, we constructed a dataset containing 78,325 music-image pairs. We demonstrate the effectiveness of the proposed mechanism on this dataset: specifically, our mechanism outperforms baseline methods by $\times 1.93 \sim 3.38$ *for the mean reciprocal rank,* $\times 2.19 \sim 3.56$ *for recall@50, and* $528 \sim 891$ *ranks for the median rank.*

1. Introduction

Can we imagine a piece of music simply by looking at its cover art? Steve and Sorger described how one of the functional parameters of cover art is to say something about the music inside [43]. Libeks *et al*. showed that cover art contains visual features that are helpful for contextualizing music [21]. Negus claimed "Different genres of music have become associated with and signify different images, which in turn connote particular attitudes, values and beliefs. [...] visual images denote particular sounds." [30]. In other words, we can indeed gain information about music just by looking at its cover art. In support of this idea, Vlad Sepetov, a designer famous for his work with Kendrick Lamar, said, "I want someone to look at the album cover and appreciate the aesthetic and image and let the artwork guide their listening experience." He continued, "... that first look at the sleeve tells you how you are going to listen to the album." [3]. Vad explained "Despite the fact that they are not in the strictest sense making sound themselves, album covers are profoundly musical. Album covers represent the music contained inside them and, even further, they mediate our listening experience. Conversely, our viewing experience is mediated by the music." [47]. In such ways, a piece of music and its cover art are designed to be closely associated with each other. The goal of this paper is to develop a method that can achieve cross-modal retrieval tasks of music and images by leveraging this association between a piece of music and its cover art, as illustrated in Figure 1.

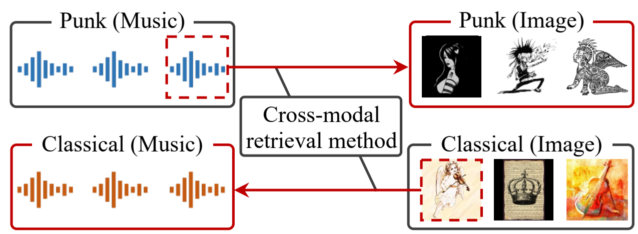

Figure 1: Conceptual design of the tasks in this study. Our objective is to develop a method for cross-modal retrieval of images that match an input piece of music (or vice versa).

Cross-modal music-image retrieval methods benefit various music information retrieval (MIR) applications. For example, these methods benefit a musician who has composed a new piece of music to find cover art for that music from a set of available images. As another example, given any new image, these methods can create a playlist of songs that match the image. Moreover, such a cross-modal retrieval method could provide insight into the latent relationship between music and images in a vast music collection.

So far, several pioneering methods related to music and images have been proposed [4, 19, 22, 28, 29, 32, 36–38, 44, 54, 58, 63, 64]. However, those methods take the approach of using metadata including tags (mood, emotion, video, *etc*.) and textual descriptions. That approach entails problems in that such metadata is not assigned to all music and images and often varies across datasets or service platforms.

In addition, it is mentioned that music and images with minor tags are difficult to retrieve [13, 48]. Accordingly, such metadata has to be consistently assigned to large amounts of data, which places a heavy burden on annotators and may require them to have technical music knowledge. Hence, in this study, we investigate a content-based music-image retrieval approach that leverages only a piece of music and its cover art without any additional metadata.

To achieve content-based music-image retrieval, we adopt a deep metric learning (DML) approach [13, 34, 45, 59], as illustrated in Figure 2. In this approach, we train two encoders that respectively embed pieces of music and images in a shared embedding space under the assumption that a pair of a piece of music and an image for the same song (*i.e.*, an original music-image pair) is positive and a pair of those for different songs is negative. Then, the encoders are trained so that the embeddings (*i.e.*, points in the shared embedding space) of a positive pair are close to each other and those of a negative pair are far from each other in the shared embedding space. Once the encoders are successfully trained, we can use them to embed a music query in the shared embedding space and retrieve images (or vice versa) that match the query according to the similarities of the embeddings in the shared embedding space.

The key to successful DML is to mine informative pairs so that a loss function returns meaningful feedback to the encoders [39, 50, 53]. The bottleneck of DML in the content-based approach is that encoders can mine a few positive instances; that is, only an original music-image pair can be a positive pair under that assumption. To overcome this bottleneck, we propose a self- and cross-modal feature embedding memory (SCFEM) mechanism that was inspired by existing feature memory mechanisms [51, 60]. The proposed mechanism stores and directly uses both the music and image feature embeddings of any previous iterations in memory. Because our mechanism enables the encoders to mine more informative positive pairs in addition to informative negative pairs from the memories than the existing mechanisms [51, 60], our mechanism is especially effective in content-based cross-modal retrieval tasks. That is, assuming that every pair between the embeddings of a piece of music and an image at a current iteration and their own stored embeddings is positive, our mechanism enables the encoders to obtain additional informative positive pairs.

To address the lack of datasets including both pieces of music and their cover art, we constructed a private dataset, called the Music Cover Art (MCA) dataset, that contains 78,325 music-image pairs (30 s audio previews for trial listening and their cover art). We then quantitatively evaluated the effectiveness of our mechanism on this dataset in terms of the mean reciprocal rank [7], recall@k, and median rank [45]. The results showed that our mechanism outperformed various baseline methods.

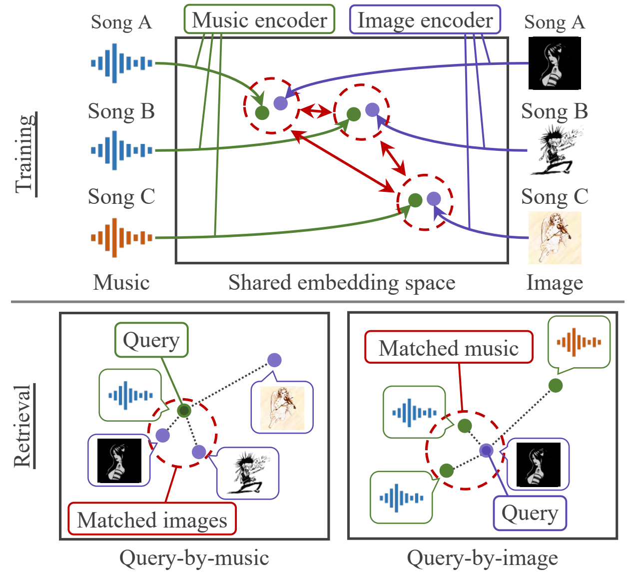

Figure 2: Overview of our approach. We train encoders so that the embeddings of the music and image for the same song are close to each other, while those for different songs are far from each other, in a shared embedding space. By calculating the similarities between embeddings in the shared embedding space, we can retrieve an image matching a given piece of music, or vice versa.

2. Related Work

2.1. Cross-Modal Music-Image Retrieval

Multimodal retrieval related to music and images has shown its potential in MIR tasks [2, 9, 27]. However, cross-modal retrieval for music and images is in the early stages of research. Mattek and Casay conducted an experiment on aesthetics in which participants were shown ten pieces of music and ten images and asked to assess their association [26]. An important aspect of that study was that it identified a cross-modal effect between music and images. In our study, we also focus on this association between music and images, especially cover art, to develop a cross-modal retrieval method for music and images.

Several studies proposed methods that used metadata including tags such as emotion and mood, and some text such as lyrics and descriptions [4, 19, 22, 28, 29, 32, 36–38, 44, 54, 58, 63, 64]. The problem is that such metadata is not necessarily assigned to all music and images. This problem may lead to the inability to perform cross-modal music-image retrieval due to the missing metadata, while a piece of music and an image are closely associated with each other. In addition, music and images may be assigned metadata that is not common to them. That is, different datasets or service platforms often assign varying kinds of metadata individually, *e.g.*, some metadata is assigned only to music (or images). Moreover, the addition of such metadata to a large amount of data places a heavy burden on anno-

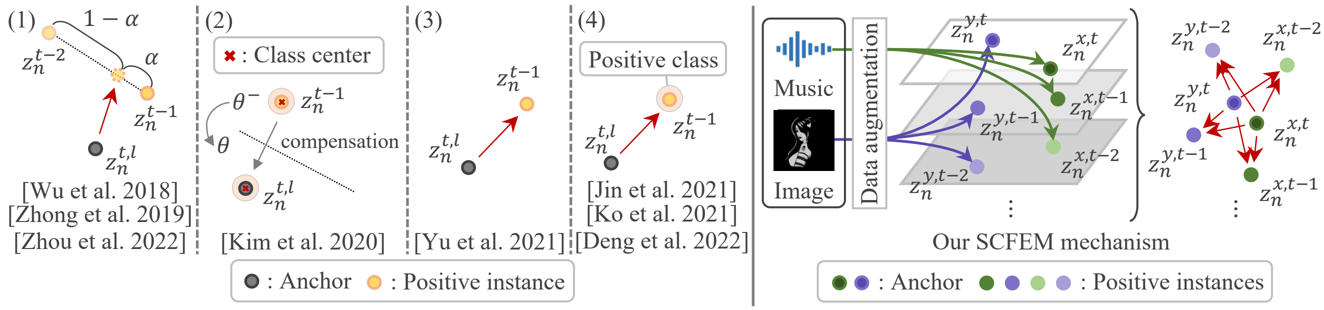

Figure 3: Explanatory diagrams of feature memory mechanisms for the case of a positive (*i.e.*, original) music-image pair. Existing mechanisms build at most one informative positive pair, whereas our proposed mechanism can build several informative positive pairs by leveraging self- and cross-modal feature embeddings. Cross-batch memory (XBM) [51] cannot mine a positive pair because it only stores embeddings of a previous iteration at a current epoch, not a past epoch.

tators and may require them to have technical knowledge of music. Accordingly, content-based cross-modal retrieval of music and images without using metadata has been proposed [13, 34, 45, 59]. Hong *et al.* proposed a soft intra-modal structure constraint in which the embeddings of instances with similar music (or images) become close to each other in a shared embedding space for content-based video-music retrieval (CBVMR) [13]. Yi *et al.* proposed a cross-modal variational autoencoder that matches the latent variables of a micro video, which includes a video, a piece of music, and short texts [59]. Prétet *et al.* investigated the effects of feature extraction modules proposed in CB-VMR [13] by replacing well-known modules with original ones [34]. Surís *et al.* proposed a transformer-based encoder that locates the embeddings of a music video computed by the contrastive language image pre-training (CLIP) [35] and the disentangled music representation learning [18] close to each other [45]. In this paper, we introduce a novel feature memory mechanism for cross-modal music-image retrieval.

2.2. Feature Memory Mechanism

A feature memory mechanism, which stores past embeddings during training and enables encoders to mine informative pairs from stored embeddings, has demonstrated its potential in a variety of computer vision tasks [11, 14, 17, 20, 49, 51, 55–57, 60, 66]. Several studies have incorporated this feature memory mechanism into cross-modal retrieval methods, *e.g.*, source code and binary code [61], an RGB image and an infrared image [23], and a food image and a cooking recipe [40]. To the best of our knowledge, the effectiveness of feature memory mechanisms has not been demonstrated in cross-modal retrieval of music and images.

As illustrated in Figure 3, the primary mechanisms for handling past embeddings are as follows: (1) updating embeddings by moving averages [55, 66, 67]; (2) compensating embeddings to adapt them to the latest network parameters [15]; (3) direct use of past embeddings [51, 60]; and (4) calculation of a representative embedding from those in the same class [8, 14, 17]. The problem is that content-based cross-modal retrieval tasks are more restrictive than other tasks in mining informative instances from a feature embedding memory. In the content-based approach, only an original music-image pair becomes a positive pair, resulting in an imbalanced number of positive and negative instances in a feature embedding memory. Therefore, it is difficult to build informative positive pairs with existing feature memory mechanisms [51, 55, 60, 66, 67], and those mechanisms cannot benefit from using classes [8, 14, 15, 17]. In contrast, our proposed mechanism can store more past embeddings than existing mechanisms can, which facilitates the building of informative positive pairs between the embeddings at a current iteration and their own stored embeddings.

3. Method

This section describes the proposed method that leverages pair-based DML. Our goal is to design two encoders that embed each piece of music and each image into a shared embedding space, and to optimize the encoders so that the embeddings of a positive music-image pair lie close to each other and those of a negative pair lie far from each other in the shared embedding space.

3.1. Problem Specification

We use a complex spectrogram of a piece of music as the input of a music encoder, following previous studies [24, 52, 65], and an RGB image as the input of an image encoder. Let $\mathbf{X} = \{\mathbf{x}_n \in \mathbb{R}^{D^{\mathbf{x}}}\}_{n=1}^{N}$ and $\mathbf{Y} = \{\mathbf{y}_n \in \mathbb{R}^{D^{\mathbf{y}}}\}_{n=1}^{N}$ be a set of complex spectrograms and a set of images corresponding to $\mathbf{X}$, respectively, where $D^{\mathbf{x}}$ is the number of dimensions of each complex spectrogram, $D^{\mathbf{y}}$ is the number of dimensions of each image, and N is the number of songs.

Next, let $\mathbf{Z}^{\mathbf{X}} = \{\mathbf{z}_n^{\mathbf{x}} \in \mathbb{R}^{D^{\mathbf{z}}}\}_{n=1}^{N}$ and $\mathbf{Z}^{\mathbf{Y}} = \{\mathbf{z}_n^{\mathbf{y}} \in \mathbb{R}^{D^{\mathbf{z}}}\}_{n=1}^{N}$ be sets of embeddings of complex spectrograms and images, respectively, where $D^{\mathbf{z}}$ is the number of dimensions of each embedding. Let $\mathcal{S}$ be a space of dimension

$D^{\mathbf{z}}$, namely, a music-image shared embedding space.

We train the music encoder $f_{\mathrm{M}}(\cdot;\boldsymbol{\theta})$ that maps $\mathbf{X}$ to $\mathbf{Z}^{\mathbf{X}}$ (*i.e.*, $\mathbf{x}_n \xmapsto{f_{\mathrm{M}}} \mathbf{z}_n^{\mathbf{x}}$) and the image encoder $f_{\mathrm{I}}(\cdot;\boldsymbol{\phi})$ that maps $\mathbf{Y}$ to $\mathbf{Z}^{\mathbf{Y}}$ (*i.e.*, $\mathbf{y}_n \xmapsto{f_{\mathrm{I}}} \mathbf{z}_n^{\mathbf{y}}$) so that the embeddings $\mathbf{z}_n^{\mathbf{x}}$ and $\mathbf{z}_n^{\mathbf{y}}$ are close in $\mathcal{S}$. Here, $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ are the parameters of the respective encoders.

3.2. Learning Framework

We first describe a basic learning framework that uses pair-based DML. Then, we introduce the key component of our SCFEM mechanism, as illustrated in Figure 4.

3.2.1 Joint Embedding Technique

A practical approach to develop a cross-modal retrieval method is to use pair-based DML such that any positive pairs lie close to each other and any negative pairs lie far from each other in a shared embedding space [13,34,45,59].

For pair-based DML, the general pair weighting (GPW) framework [50] provided the GPW formulation $\mathcal{F}(\boldsymbol{B})$ for analyzing a pair-based loss function $\mathcal{L}(\boldsymbol{B})$ as follows:

$$\begin{aligned}\mathcal{F}(\boldsymbol{B}) &= \frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{m}\left.\frac{\partial\mathcal{L}(\boldsymbol{B})}{\partial B_{ij}}\right|_{l} B_{ij} \\ &= \frac{1}{m}\sum_{i=1}^{m}\left(\sum_{(\mathbf{x}_i,\mathbf{y}_j)\in\mathcal{N}} w_{ij}^{B} B_{ij} - \sum_{(\mathbf{x}_i,\mathbf{y}_j)\in\mathcal{P}} w_{ij}^{B} B_{ij}\right). \quad (1)\end{aligned}$$

Here, m is a mini-batch size; $\mathcal{P}$ and $\mathcal{N}$ are a set of positive pairs and a set of negative pairs, respectively; $w_{ij}^{B} = \left|\left.\frac{\partial\mathcal{L}(\boldsymbol{B})}{\partial B_{ij}}\right|_{l}\right|$ is a weight at the l-th iteration; and $\boldsymbol{B}$ is a similarity matrix whose element (i,j) is defined as the cosine similarity between $\mathbf{z}_i^{\mathbf{x}}$ and $\mathbf{z}_j^{\mathbf{y}}$ (*i.e.*, $B_{ij} = \mathrm{sim}(\mathbf{z}_i^{\mathbf{x}}, \mathbf{z}_j^{\mathbf{y}}) = {\mathbf{z}_i^{\mathbf{x}}}^{\mathsf{T}}\mathbf{z}_j^{\mathbf{y}}/|\mathbf{z}_i^{\mathbf{x}}||\mathbf{z}_j^{\mathbf{y}}|$). Eq. (1) indicates that it is important to appropriately design the mini-batch size m that controls the number of possible pairs, the weight w_{ij}^{B} that is assigned to B_{ij}, and the sets of pairs $\mathcal{P}$ and $\mathcal{N}$, which should consist of informative pairs for training.

For pair-based cross-modal DML, we can build two types of pairs [13,34,45,59]: one in which a piece of music is used as an anchor, and another in which an image is used as an anchor. Thus, Eq. (1) can be rewritten as follows:

$$\begin{aligned}\mathcal{F}(\boldsymbol{B}) = \frac{1}{m}\sum_{i=1}^{m}&\left(\sum_{(\mathbf{x}_i,\mathbf{y}_j)\in\mathcal{N}} w_{ij}^{B} B_{ij} + \sum_{(\mathbf{y}_i,\mathbf{x}_j)\in\mathcal{N}} \hat{w}_{ij}^{B} \hat{B}_{ij}\right) \\ - \frac{1}{m}\sum_{i=1}^{m}&\left(\sum_{(\mathbf{x}_i,\mathbf{y}_j)\in\mathcal{P}} w_{ij}^{B} B_{ij} + \sum_{(\mathbf{y}_i,\mathbf{x}_j)\in\mathcal{P}} \hat{w}_{ij}^{B} \hat{B}_{ij}\right), \quad (2)\end{aligned}$$

where $\hat{w}_{ij}^{B} = \left|\left.\frac{\partial\mathcal{L}(\boldsymbol{B})}{\partial \hat{B}_{ij}}\right|_{l}\right|$ and $\hat{B}_{ij} = \mathrm{sim}(\mathbf{z}_i^{\mathbf{y}}, \mathbf{z}_j^{\mathbf{x}})$.

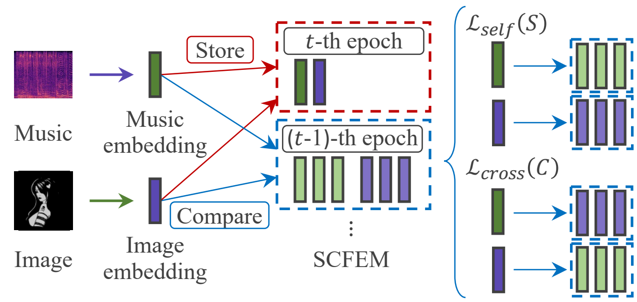

Figure 4: Explanatory diagrams of our proposed mechanism, which trains encoders by comparing each anchor with embeddings in memory. The embeddings at the current iteration are stored in the memory. The mechanism enables us to define loss functions using both self- and cross-modal feature embedding memory.

The policy of our problem specification is the same as that of existing content-based cross-modal retrieval methods [13, 34, 45, 59]. The difference is that we address the tasks in the music-image domain, whereas theirs are in the music-video domain. For this problem specification, using contrastive learning [5, 11, 31, 42] is an effective approach [45]. Here, we use a contrastive loss function called InfoNCE [31] as follows:

$$\begin{aligned}\mathcal{L}_{batch}(\boldsymbol{B}) = &-\frac{1}{m}\sum_{i=1}^{m}\log\frac{e^{B_{i+}/\tau}}{\sum_{j=1}^{m} e^{B_{ij}/\tau}} \\ &-\frac{1}{m}\sum_{i=1}^{m}\log\frac{e^{\hat{B}_{i+}/\tau}}{\sum_{j=1}^{m} e^{\hat{B}_{ij}/\tau}}, \quad (3)\end{aligned}$$

where τ is a hyperparameter called temperature scaling that controls the scale of the loss function and $+$ indicates a positive instance of an anchor. Each term in the r.h.s. of Eq. (3) indicates each type of pairs considered in Eq. (2). The weight w_{ij}^{B} is derived from Eqs. (2) and (3) as follows:

$$w_{ij}^{B} = \begin{cases}\frac{1}{\tau} - \chi_{ij}^{B} & ((\mathbf{x}_i,\mathbf{y}_j)\in\mathcal{P}), \\ \chi_{ij}^{B} & ((\mathbf{x}_i,\mathbf{y}_j)\in\mathcal{N}),\end{cases} \quad (4)$$

where $\chi_{ij}^{B} = e^{B_{ij}/\tau}/\{\tau(e^{B_{i+}/\tau} + \sum_{(\mathbf{x}_i,\mathbf{y}_j)\in\mathcal{N}} e^{B_{ij}/\tau})\}$. The weight $\hat{w}_{ij}^{B}$ can be derived as well as w_{ij}^{B}.

We then estimate optimal parameters $\boldsymbol{\theta}^*$ and $\boldsymbol{\phi}^*$ by minimizing the loss function $\mathcal{L}_{batch}$ as follows:

$$\boldsymbol{\theta}^*, \boldsymbol{\phi}^* = \mathop{\arg\min}_{\boldsymbol{\theta},\boldsymbol{\phi}} \mathcal{L}_{batch}. \quad (5)$$

3.2.2 Self- and Cross-Modal Feature Embedding Memory

Inspired by the "slow drift" phenomenon [51], we propose a new mechanism called self- and cross-modal feature embedding memory (SCFEM). This mechanism can be seamlessly integrated into a pair-based DML framework as a module, and can perform with a small amount of computational resources even though our mechanism can handle a sufficiently large number of instances larger than the mini-batch size at each training iteration.

Let $\boldsymbol{M}^{\mathbf{x}}, \boldsymbol{M}^{\mathbf{y}} \in \mathbb{R}^{N \times D^{\mathbf{z}} \times E}$ be a music feature embedding memory and an image feature embedding memory, respectively, where E is the number of epochs to be stored in the feature embedding memories. Our mechanism first requires initialization of feature embedding memories $\boldsymbol{M}^{\mathbf{x}}$ and $\boldsymbol{M}^{\mathbf{y}}$ at the beginning of training. Our mechanism can be triggered once the encoders are warmed up (*i.e.*, training has stabilized at a local optimal parameters of the encoders). At each iteration, embeddings are stored in the feature embedding memories. When the number of stored embeddings exceeds the size of feature embedding memories, the earliest embeddings stored in the feature embedding memories are replaced with the embeddings at the current iteration.

Here, the important aspect of the proposed mechanism is that we can define two loss functions—one using a self-modal feature embedding memory, and the other using a cross-modal feature embedding memory—because of the availability of both the music and image feature embedding memories. That is, the proposed mechanism enables the encoders to mine informative pairs from both the music and image feature embedding memories. Let $\mathcal{L}_{self}$ and $\mathcal{L}_{cross}$ be loss functions using self- and cross-modal feature embedding memory, respectively. As in Eq. (3), the loss function $\mathcal{L}_{self}$ can be written as follows:

$$\begin{aligned}\mathcal{L}_{self}(\boldsymbol{S}) = &- \frac{1}{m}\sum_{i=1}^{m}\sum_{e=0}^{E-1}\log\frac{w_e e^{S^{\mathbf{x}}_{i+}/\tau}}{\sum_{j=1}^{N} e^{S^{\mathbf{x}}_{ij}/\tau}} \\ &- \frac{1}{m}\sum_{i=1}^{m}\sum_{e=0}^{E-1}\log\frac{w_e e^{S^{\mathbf{y}}_{i+}/\tau}}{\sum_{j=1}^{N} e^{S^{\mathbf{y}}_{ij}/\tau}},\end{aligned} \tag{6}$$

where $\boldsymbol{S}$ is a similarity matrix whose element (i, j) is defined as the cosine similarity between an instance of a mini-batch and an instance stored in the self-modal feature embedding memory (*i.e.*, $S^{\mathbf{x}}_{ij} = \mathrm{sim}(\mathbf{z}^{\mathbf{x},t}_i, \mathbf{z}^{\mathbf{x},t-e}_j \in \boldsymbol{M}^{\mathbf{x}})$ and $S^{\mathbf{y}}_{ij} = \mathrm{sim}(\mathbf{z}^{\mathbf{y},t}_i, \mathbf{z}^{\mathbf{y},t-e}_j \in \boldsymbol{M}^{\mathbf{y}})$), and $\{w_e\}_{e=0}^{E-1}$ is a set of weights. Similarly to $\mathcal{L}_{self}$, the loss function $\mathcal{L}_{cross}$ can be written as follows:

$$\begin{aligned}\mathcal{L}_{cross}(\boldsymbol{C}) = &- \frac{1}{m}\sum_{i=1}^{m}\sum_{e=0}^{E-1}\log\frac{w_e e^{C_{i+}/\tau}}{\sum_{j=1}^{N} e^{C_{ij}/\tau}} \\ &- \frac{1}{m}\sum_{i=1}^{m}\sum_{e=0}^{E-1}\log\frac{w_e e^{\hat{C}_{i+}/\tau}}{\sum_{j=1}^{N} e^{\hat{C}_{ij}/\tau}}.\end{aligned} \tag{7}$$

where $\boldsymbol{C}$ is a similarity matrix whose element (i, j) is defined as the cosine similarity between an instance of a mini-batch and an instance stored in the cross-modal feature embedding memory (*i.e.*, $C_{ij} = \mathrm{sim}(\mathbf{z}^{\mathbf{x},t}_i, \mathbf{z}^{\mathbf{y},t-e}_j \in \boldsymbol{M}^{\mathbf{y}})$ and $\hat{C}_{ij} = \mathrm{sim}(\mathbf{z}^{\mathbf{y},t}_i, \mathbf{z}^{\mathbf{x},t-e}_j \in \boldsymbol{M}^{\mathbf{x}})$). See the supplementary material for a detailed analysis of loss functions $\mathcal{L}_{self}$ and $\mathcal{L}_{cross}$ using the GPW formulation.

Finally, by including both loss functions, $\mathcal{L}_{self}$ and $\mathcal{L}_{cross}$, in Eq. (5), we can thus estimate the optimal parameters $\boldsymbol{\theta}^*$ and $\boldsymbol{\phi}^*$ as follow:

$$\boldsymbol{\theta}^*, \boldsymbol{\phi}^* = \operatorname*{arg\,min}_{\boldsymbol{\theta},\boldsymbol{\phi}} \left(\mathcal{L}_{batch} + \lambda_{self}\mathcal{L}_{self} + \lambda_{cross}\mathcal{L}_{cross}\right), \tag{8}$$

where λ_{self} and λ_{cross} are weights that balance the loss functions.

3.3. Cross-Modal Music-Image Retrieval

Once the training of the encoders has been completed, we can estimate the similarity of a given pair of a piece of music and a cover art image as follow. First, we calculate the complex spectrogram of a given piece of music. We then use the trained encoders to obtain embeddings of the complex spectrogram and the cover art image. Finally, we compute the similarity between the obtained embeddings. A high similarity indicates that a given pair is matched.

4. Experiments and Results

This section describes comparison experiments to evaluate the effectiveness of our mechanism. To quantitatively evaluate the performance of each method, we set up two tasks: query-by-music, in which a piece of music is used as a query to retrieve a corresponding image; and query-by-image, in which an image is used as a query to retrieve a corresponding piece of music. This section also describes a qualitative analysis of the obtained embeddings.

4.1. Experimental Setup

4.1.1 Dataset

We constructed the private MCA dataset that contains pairs of a music excerpt (approximately 30 s audio signal with 44.1 kHz sampling rate) for trial listening and its cover art (square shaped RGB image). These music excerpts (typically, representative music sections) had already been cropped on an Internet music service from which the excerpts were crawled as is often done by other studies [1, 45, 59]. The corresponding cover art images were crawled at the same time. We collected songs so that each song was associated with a different cover art image and cover art image was associated with a different song (*i.e.*, one-to-one correspondence between music and image). This dataset contains 78,325 songs by 40,151 artists and encompasses a variety of music genres (over 250, according to the service). We randomly split the dataset into training, validation, and test sets with an eight-one-one ratio (*i.e.*, training set: 62,659 songs; validation set: 7,833 songs; test set: 7,833 songs).

	Query-by-music				Query-by-image			
	MRR	R@50	R@100	Median rank	MRR	R@50	R@100	Median rank
Random	1.22×10^{-3}	0.64	1.28	3917	1.22×10^{-3}	0.64	1.28	3917
CBVMR [13]	1.34×10^{-3}	0.75	1.52	3686	1.27×10^{-3}	0.61	1.39	3656
Baseline (HRFormer [62])	3.37×10^{-3}	2.09	4.05	1957	3.42×10^{-3}	2.08	4.06	1926
w/ Data Augmentation	3.82×10^{-3}	2.84	5.57	1614	3.81×10^{-3}	2.56	5.22	1626
+ w/ XBM [51]	4.23×10^{-3}	2.78	5.86	1594	5.04×10^{-3}	3.22	6.09	1600
+ w/ SCFEM (Ours)	$\mathbf{1.14 \times 10^{-2}}$	**7.45**	**12.3**	**1066**	$\mathbf{9.75 \times 10^{-3}}$	**7.06**	**11.8**	**1059**

Table 1: Results for MRR, R@k, and the median rank on the test set of the MCA dataset in the query-by-music and query-by-image settings, with k set to 50 and 100, respectively.

4.1.2 Implementation Details

Music Representation: The complex spectrogram was calculated by the short-time Fourier transform (STFT) [10] using nnAudio [6] with a Hann window, frequency bins F of 1,025, and a stride size of 512. Then, the complex spectrogram was cropped so that the shape of the cropped complex spectrogram was $2 \times F \times 256$ (*i.e.*, a music audio signal with a frame length of approximately 3 s). The music encoder embeds the cropped complex spectrogram into the 256-dimensional shared embedding space. While training the music encoder, we randomly cropped the complex spectrogram for data augmentation. For the test, we used the averaged value of the embeddings of the cropped complex spectrograms for each piece of music, where we iteratively cropped the complex spectrogram from the beginning of the music audio signal with 50% overlapping.

Image Representation: The image was resized to 256 px $\times$ 256 px. The image encoder embeds the resized image into the 256-dimensional shared embedding space. While training the image encoder, an affine transformation including random rotation ($[-25°, 25°]$), random translation ($[0.15, 0.15]$), and random scaling ($[0.75, 1.25]$) was applied to all the images for data augmentation.

Encoder Architecture: We used HRFormer [62] as a backbone network. The final layer of the backbone network was set as an embedding layer instead of a classifier.

Training Options: We trained the encoders from scratch and warmed them up with over 50k iterations. Our implementation was based on PyTorch [33]. We used the Adam optimizer [16] with a learning rate of 1.0×10^{-4}. We used eight NVIDIA A100 40-GB PCIe GPU Accelerators for three days for training. We empirically set the weights ($\lambda_{self} = 0.3, \lambda_{cross} = 0.2$) regarding the loss functions so that the values of each loss function would be approximately equal. We also set the temperature-scaling value that was originally used in MOCO [31] (*i.e.*, $\tau = 0.07$).

4.1.3 Ranking-based Evaluation Metrics

We used three standard evaluation metrics in cross-modal tasks for the comparison experiments: the mean reciprocal rank (MRR) [7], the recall@k (R@k), and the median rank [45].

4.2. Conditions

To demonstrate the effectiveness of the proposed mechanism, we compared it with the following baseline methods.

- **Baseline:** HRFormer [62] as a backbone network for each encoder without any data augmentation or feature memory mechanisms.
- **Baseline w/ Data Augmentation:** HRFormer as a backbone network for each encoder with data augmentation and no feature memory mechanisms.
- **Baseline + w/ XBM:** HRFormer as a backbone network for each encoder with data augmentation and a cross-batch memory (XBM) mechanism [51]. In this study, XBM is the same as the proposed mechanism when $E = 1$. This baseline is also comparable to the cross-epoch learning [60], although their method uses negative instances stored in the memory at one previous epoch.
- **Baseline + w/ SCFEM (ours):** HRFormer as a backbone network for each encoder with data augmentation and our proposed SCFEM. We here set $E = 2$ and $w_0 = w_1 = 1.0$.

In addition, we include the results of the following methods here for reference.

- **Random:** We used random estimation.
- **CBVMR:** We tested CBVMR [13], but it differs from our study in terms of the input representations because it focuses on cross-modal retrieval for music and video (not image). Instead of video-level features, we directly used frame-level features with a whitened principal component analysis described in their paper.

4.3. Results

Table 1 lists the MRR, R@k, and median rank results in the query-by-music and query-by-image settings. Our proposed mechanism outperformed the baseline methods by $\times 2.70 \sim 3.38$ for the MRR, $\times 2.62 \sim 3.56$ for R@50,

	Query-by-music				Query-by-image			
	MRR	R@50	R@100	Median rank	MRR	R@50	R@100	Median rank
$E = 1$ (XBM [51])	4.23×10^{-3}	2.78	5.86	1594	5.04×10^{-3}	3.22	6.09	1600
$E = 2, w_1 = 1.0$	1.14×10^{-2}	7.45	12.3	1066	9.75×10^{-3}	7.06	11.8	1059
$E = 3, w_1 = w_2 = 0.5$	1.10×10^{-2}	8.00	12.9	1014	9.49×10^{-3}	6.92	12.2	1002
$E = 3, w_1 = 0.6, w_2 = 0.4$	1.13×10^{-2}	**8.04**	12.9	1034	1.05×10^{-2}	**7.50**	12.3	1010
$E = 3, w_1 = 0.7, w_2 = 0.3$	1.11×10^{-2}	7.53	12.7	1014	$\mathbf{1.09} \times 10^{-2}$	7.49	12.0	**982**
$E = 3, w_1 = 0.8, w_2 = 0.2$	$\mathbf{1.26} \times 10^{-2}$	7.78	**13.0**	1024	$\mathbf{1.09} \times 10^{-2}$	7.34	**12.5**	1022
$E = 3, w_1 = 0.9, w_2 = 0.1$	1.10×10^{-2}	7.91	12.9	**1009**	1.07×10^{-2}	7.47	12.2	1010

Table 2: Comparison of memory sizes and weights.

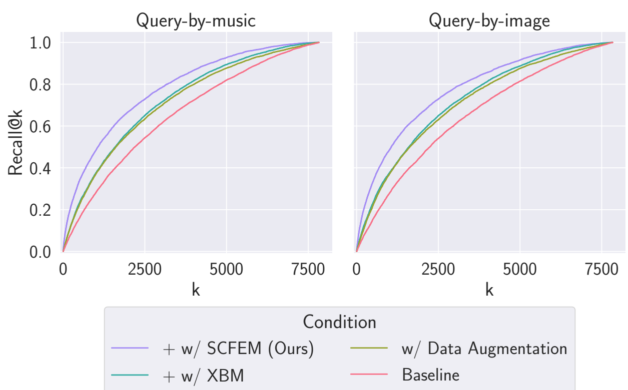

Figure 5: Empirical cumulative distribution functions (CDFs) for k in the query-by-music and the query-by-image settings.

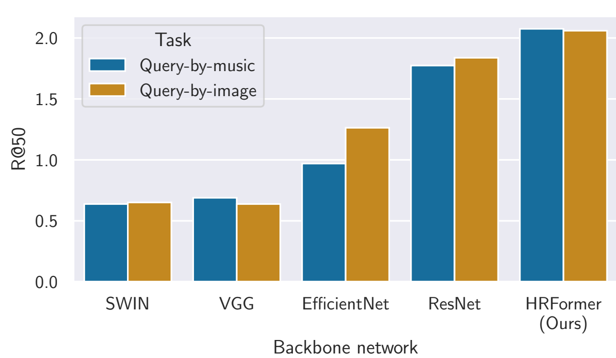

Figure 6: Comparison of backbone networks.

and 528 $\sim$ 891 ranks in the query-by-music setting; and by $\times 1.93 \sim 2.85$ for the MRR, $\times 2.19 \sim 3.39$ for R@50, and 541 $\sim$ 867 ranks in the query-by-image setting. Figure 5 shows empirical cumulative distribution functions (CDFs) with respect to k in both settings. These CDFs illustrate the advantage of our SCFEM mechanism for almost all R@k levels. This result demonstrates that our mechanism, which can mine more informative instances from the feature memories, was effective in the retrieval tasks of this study.

Moreover, the baseline method with data augmentation was superior to that without data augmentation. This result indicates that the data augmentation we used was effective for training, whereas existing content-based cross-modal retrieval methods [13, 34, 45, 59] directly used music and image features as training data without data augmentation.

4.4. Ablation and Comparative Study

We provide ablation and comparative study to verify the effectiveness of each component in our SCFEM mechanism and the necessity of warmed-up encoders.

4.4.1 Backbone Network

Selection of the backbone network has a large impact on performance. To investigate the impact, we compared several well-known neural network models as backbones, including CNN-based models [12, 41, 46] and Transformer-based models [25, 62]. We used $\tau = 1.0$ in this comparison experiment. Figure 6 shows the results, which confirm the appropriateness of using HRFormer [62] as the backbone in this paper.

4.4.2 Memory Size

Since our SCFEM mechanism can store music and image embeddings of more previous iterations in memory and leverage all of them to obtain more positive instances, we compared the performance with different memory sizes and weights ($w_0 = 1.0$ is fixed for all conditions). Although $E = 2$ and $w_0 = w_1 = 1.0$ was used in Table 1, the results listed in Table 2 indicate that an increase in memory size can further improve the performance. Note that it is necessary to set appropriate weights when increasing the memory size. We leave the investigation of their optimal setting for future work.

4.4.3 Embedding Size

To investigate the effect of the number of dimensions of the shared embedding space, we compared the R@50 performances with $D^{\mathbf{z}} = 64, 128, 256$, and 512. The results

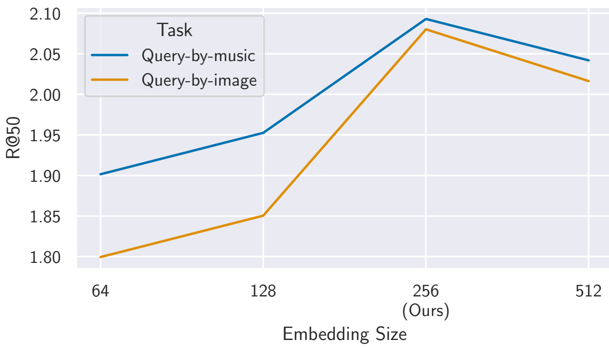

Figure 7: Comparison of embedding sizes.

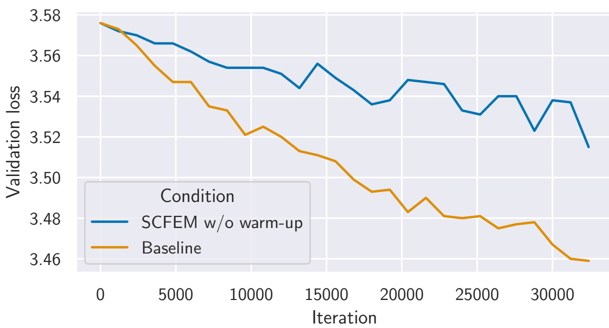

Figure 8: Validation losses for the proposed mechanism without warm-up and Baseline (HRFormer [62]).

shown in Figure 7 confirm that $D^{\mathbf{z}} = 256$, which was used for Table 1, is the best choice.

4.4.4 Necessity of Warmed-Up Encoders

Since the parameters of encoders are largely updated at the initial stage of training, their embeddings are changed too much during iterations and are not expected to be informative instances in the memory. Our SCFEM mechanism is therefore applied only after the encoders are warmed up as described in Section 3.2. In Figure 8, we compared validation losses for the proposed mechanism without the warm-up and the baseline method, which confirms that adverse effects of performance deterioration occur if our mechanism is applied too early in the training (*i.e.*, the warm-up is necessary).

4.5. Qualitative Analysis

A qualitative analysis was also conducted to further investigate the nature of the obtained embeddings. We applied principal component analysis (PCA) on the embeddings of the music and images for 686 songs in total categorized as Metal, Jazz, Classical, Electronic, and Punk in the test set (note that metadata including these category tags are not used at all in our training). Figure 9 shows that embeddings for songs of the same category are relatively close to each

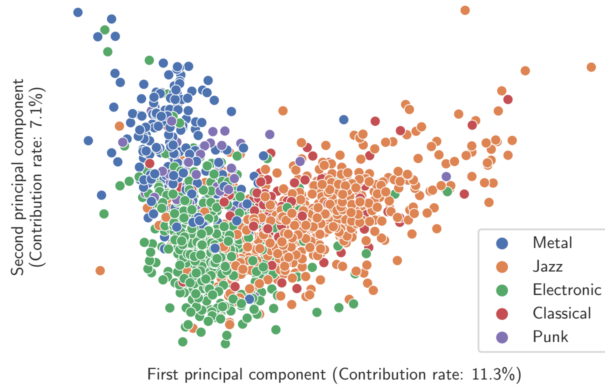

Figure 9: Principal component analysis for both music and image embeddings.

other in the shared embedding space. Interestingly, the embeddings of Metal and Punk songs are close to each other in the shared embedding space because of their similarities in pieces of music and images. This result supports the association between a piece of music and its cover art described in Section 1.

5. Conclusion

In this study of content-based music-image retrieval, we proposed a mechanism called self- and cross-modal feature embedding memory (SCFEM), which can be seamlessly integrated into a pair-based DML framework. The contributions of this paper can be summarized as follows. First, the proposed mechanism can store the embeddings of any previous iterations in order to mine informative pairs from the feature memories. This approach leverages the power of the feature embedding memory mechanism for music-image retrieval tasks. Second, our comparison experiments using ranking-based evaluation metrics (*i.e.*, the mean reciprocal rank, recall@k, and median rank) demonstrated that our mechanism outperformed the baseline methods. We also demonstrated that an increase in memory size improved the performance. Third, the qualitative analysis reveals that music and images similar in style are close to each other in the shared embedding space.

The proposed mechanism can also be applied to hard mining problems in not only MIR tasks but also other computer vision tasks. We believe that this proposed mechanism opens up the possibilities of achieving a broad range of cross-modal tasks.

Acknowledgements

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

References

[1] Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. Youtube-8m: A large-scale video classification benchmark. *arXiv preprint arXiv:1609.08675*, 2016.

[2] Eric Brochu, Nando De Freitas, and Kejie Bao. The sound of an album cover: Probabilistic multimedia and information retrieval. In *Proceedings of the International Workshop on Artificial Intelligence and Statistics*, pages 49–56, 2003.

[3] Jim Carroll. The art of the sleeve: every album cover tells a story, 2016. Available: `https://www.irishtimes.com/culture/music/the-art-of-the-sleeve-every-album-cover-tells-a-story-1.2821084` (accessed July 7, 2022).

[4] Jiansong Chao, Haofen Wang, Wenlei Zhou, Weinan Zhang, and Yong Yu. TuneSensor: A semantic-driven music recommendation service for digital photo albums. In *Proceedings of the International Semantic Web Conference (ISWC)*, 2011.

[5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 1597–1607, 2020.

[6] Kin Wai Cheuk, Hans Anderson, Kat Agres, and Dorien Herremans. nnAudio: An on-the-fly GPU audio to spectrogram conversion toolbox using 1D convolutional neural networks. *IEEE Access*, 8:161981–162003, 2020.

[7] Nick Craswell. *Mean Reciprocal Rank*, page 1703. Springer US, 2009.

[8] Zelu Deng, Yujie Zhong, Sheng Guo, and Weilin Huang. InsCLR: Improving instance retrieval with self-supervision. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 516–524, 2022.

[9] Peter Dunker, Stefanie Nowak, André Begau, and Cornelia Lanz. Content-based mood classification for photos and music: A generic multi-modal classification framework and evaluation approach. In *Proceedings of the ACM International Conference on Multimedia Information Retrieval*, pages 97–104, 2008.

[10] D. Gabor. Theory of communication. *Journal of the Institution of Electrical Engineers*, 94(73), 1947.

[11] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9729–9738, 2020.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[13] Sungeun Hong, Woobin Im, and Hyun S Yang. CBVMR: Content-based video-music retrieval using soft intra-modal structure constraint. In *Proceedings of the ACM International Conference on Multimedia Retrieval (ICMR)*, pages 353–361, 2018.

[14] Zhenchao Jin, Tao Gong, Dongdong Yu, Qi Chu, Jian Wang, Changhu Wang, and Jie Shao. Mining contextual information beyond image for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7231–7241, 2021.

[15] Yonghyun Kim, Wonpyo Park, and Jongju Shin. BroadFace: Looking at tens of thousands of people at once for face recognition. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 536–552, 2020.

[16] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proceedings of the International Conference for Learning Representations (ICLR)*, pages 1–13, 2015.

[17] Byungsoo Ko, Geonmo Gu, and Han-Gyu Kim. Learning with memory-based virtual classes for deep metric learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11792–11801, 2021.

[18] Jongpil Lee, Nicholas J Bryan, Justin Salamon, Zeyu Jin, and Juhan Nam. Metric learning vs classification for disentangled music representation learning. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 439–445, 2020.

[19] Bochen Li and Aparna Kumar. Query by video: Cross-modal music retrieval. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 604–611, 2019.

[20] Suichan Li, Dapeng Chen, Bin Liu, Nenghai Yu, and Rui Zhao. Memory-based neighbourhood embedding for visual recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6102–6111, 2019.

[21] Janis Libeks and Douglas Turnbull. You can judge an artist by an album cover: Using images for music annotation. *IEEE MultiMedia*, 18(4):30–37, 2011.

[22] Chien-Liang Liu and Ying-Chuan Chen. Background music recommendation based on latent factors and moods. *Knowledge-Based Systems*, 159:158–170, 2018.

[23] Jialun Liu, Yifan Sun, Feng Zhu, Hongbin Pei, Yi Yang, and Wenhui Li. Learning memory-augmented unidirectional metrics for cross-modality person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19366–19375, 2022.

[24] Yun Liu, Hui Zhang, Xueliang Zhang, and Linju Yang. Supervised speech enhancement with real spectrum approximation. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5746–5750, 2019.

[25] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin Transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10012–10022, 2021.

[26] Alison Mattek and Michael A Casey. Cross-modal aesthetics from a feature extraction perspective: A pilot study. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 585–590, 2011.

[27] Rudolf Mayer. Analysing the similarity of album art with self-organising maps. In *Proceedings of the International Workshop on Self-Organizing Maps*, pages 357–366, 2011.

[28] Pratik Mazumder, Pravendra Singh, Kranti Kumar Parida, and Vinay P Namboodiri. Avgzslnet: Audio-visual generalized zero-shot learning by reconstructing label features from multi-modal embeddings. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 3090–3099, 2021.

[29] Otniel-Bogdan Mercea, Lukas Riesch, A Koepke, and Zeynep Akata. Audio-visual generalised zero-shot learning with cross-modal attention and language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10553–10563, 2022.

[30] Keith Negus. *Producing pop: Culture and conflict in the popular music industry*. Edward Arnold, 2011.

[31] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[32] Kranti Parida, Neeraj Matiyali, Tanaya Guha, and Gaurav Sharma. Coordinated joint multimodal embeddings for generalized audio-visual zero-shot classification and retrieval of videos. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 3251–3260, 2020.

[33] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. PyTorch: An imperative style, high-performance deep learning library. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, pages 8024–8035, 2019.

[34] Laure Prétet, Gael Richard, and Geoffroy Peeters. Cross-modal music-video recommendation: A study of design choices. In *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*, pages 1–9, 2021.

[35] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 8748–8763, 2021.

[36] Shoto Sasaki, Tatsunori Hirai, Hayato Ohya, and Shigeo Morishima. Affective music recommendation system reflecting the mood of input image. In *Proceedings of the International Conference on Culture and Computing (ICCC)*, pages 153–154, 2013.

[37] Shoto Sasaki, Tatsunori Hirai, Hayato Ohya, and Shigeo Morishima. Affective music recommendation system based on the mood of input video. In *Proceedings of the International Conference on Multimedia Modeling (MMM)*, pages 299–302, 2015.

[38] Alexander Schindler and Andreas Rauber. Harnessing music-related visual stereotypes for music information retrieval. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 8(2):1–21, 2016.

[39] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 815–823, 2015.

[40] Mustafa Shukor, Guillaume Couairon, Asya Grechka, and Matthieu Cord. Transformer decoders with multimodal regularization for cross-modal food retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4567–4578, 2022.

[41] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *Proceedings of the International Conference on Learning Representations (ICLR)*, pages 1–14, 2015.

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

[43] Jones Steve and Martin Sorger. Covering music: A brief history and analysis of album cover design. *Journal of Popular Music Studies*, 11(1):68–102, 1999.

[44] Didac Surís, Amanda Duarte, Amaia Salvador, Jordi Torres, and Xavier Giró-i Nieto. Cross-modal embeddings for video and audio retrieval. In *Proceedings of the European Conference on Computer Vision Workshops (ECCVW)*, 2018.

[45] Dídac Surís, Carl Vondrick, Bryan Russell, and Justin Salamon. It's time for artistic correspondence in music and video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10564–10574, 2022.

[46] Mingxing Tan and Quoc Le. EfficientNet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 6105–6114, 2019.

[47] Mikkel Vad. The album cover. *Journal of Popular Music Studies*, 33(3):11–15, 2021.

[48] Aaron Van den Oord, Sander Dieleman, and Benjamin Schrauwen. Deep content-based music recommendation. In *Proceedings of the Advances in Neural Information Processing Systems (NIPS)*, pages 2643–2651, 2013.

[49] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. In *Proceedings of the Advances in Neural Information Processing Systems (NIPS)*, pages 3630–3638, 2016.

[50] Xun Wang, Xintong Han, Weilin Huang, Dengke Dong, and Matthew R Scott. Multi-similarity loss with general pair weighting for deep metric learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5022–5030, 2019.

[51] Xun Wang, Haozhi Zhang, Weilin Huang, and Matthew R Scott. Cross-batch memory for embedding learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6388–6397, 2020.

[52] Yuxuan Wang and DeLiang Wang. A deep neural network for time-domain signal reconstruction. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4390–4394, 2015.

[53] Chao-Yuan Wu, R Manmatha, Alexander J Smola, and Philipp Krahenbuhl. Sampling matters in deep embedding

learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2840–2848, 2017.

[54] Xixuan Wu, Yu Qiao, Xiaogang Wang, and Xiaoou Tang. Bridging music and image via cross-modal ranking analysis. *IEEE Transactions on Multimedia (TOM)*, 18(7):1305–1318, 2016.

[55] Zhirong Wu, Alexei A Efros, and Stella X Yu. Improving generalization via scalable neighborhood component analysis. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 685–701, 2018.

[56] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3733–3742, 2018.

[57] Tong Xiao, Shuang Li, Bochao Wang, Liang Lin, and Xiaogang Wang. Joint detection and identification feature learning for person search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3415–3424, 2017.

[58] Baixi Xing, Kejun Zhang, Lekai Zhang, Xinda Wu, Jian Dou, and Shouqian Sun. Image–music synesthesia-aware learning based on emotional similarity recognition. *IEEE Access*, 7:136378–136390, 2019.

[59] Jing Yi, Yaochen Zhu, Jiayi Xie, and Zhenzhong Chen. Cross-modal variational auto-encoder for content-based micro-video background music recommendation. *IEEE Transactions on Multimedia (TOM)*, 2021.

[60] Shenghao Yu, Chong Wang, Qiaomei Mao, Yuqi Li, and Jiafei Wu. Cross-epoch learning for weakly supervised anomaly detection in surveillance videos. *IEEE Signal Processing Letters*, 28:2137–2141, 2021.

[61] Zeping Yu, Wenxin Zheng, Jiaqi Wang, Qiyi Tang, Sen Nie, and Shi Wu. CodeCMR: Cross-modal retrieval for function-level binary source code matching. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, pages 3872–3883, 2020.

[62] Yuhui Yuan, Rao Fu, Lang Huang, Weihong Lin, Chao Zhang, Xilin Chen, and Jingdong Wang. HRFormer: High-resolution transformer for dense prediction. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[63] D. Zeng, Y. Yu, and K. Oyama. Audio-visual embedding for cross-modal music video retrieval through supervised deep CCA. In *Proceedings of the IEEE International Symposium on Multimedia (ISM)*, pages 143–150, 2018.

[64] Sicheng Zhao, Yaxian Li, Xingxu Yao, Weizhi Nie, Pengfei Xu, Jufeng Yang, and Kurt Keutzer. Emotion-based end-to-end matching between image and music in valence-arousal space. In *Proceedings of the ACM International Conference on Multimedia*, pages 2945–2954, 2020.

[65] Naijun Zheng and Xiao-Lei Zhang. Phase-aware speech enhancement based on deep neural networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, 27(1):63–76, 2018.

[66] Zhun Zhong, Liang Zheng, Zhiming Luo, Shaozi Li, and Yi Yang. Invariance matters: Exemplar memory for domain adaptive person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 598–607, 2019.

[67] Tianfei Zhou, Meijie Zhang, Fang Zhao, and Jianwu Li. Regional semantic contrast and aggregation for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4299–4309, 2022.