Emotion-Based End-to-End Matching Between Image and Music in Valence-Arousal Space

Sicheng Zhao*†
schzhao@gmail.com
University of California, Berkeley

Yaxian Li†
liyaxian@ruc.edu.cn
Renmin University of China

Xingxu Yao†
yxx_hbgd@163.com
Nankai University, Didi Chuxing

Weizhi Nie
weizhinie@tju.edu.cn
Tianjin University

Pengfei Xu
xupengfeipf@didiglobal.com
Didi Chuxing

Jufeng Yang
yangjufeng@nankai.edu.cn
Nankai University

Kurt Keutzer
keutzer@berkeley.edu
University of California, Berkeley

ABSTRACT

Both images and music can convey rich semantics and are widely used to induce specific emotions. Matching images and music with similar emotions might help to make emotion perceptions more vivid and stronger. Existing emotion-based image and music matching methods either employ limited categorical emotion states which cannot well reflect the complexity and subtlety of emotions, or train the matching model using an impractical multi-stage pipeline. In this paper, we study end-to-end matching between image and music based on emotions in the continuous valence-arousal (VA) space. First, we construct a large-scale dataset, termed Image-Music-Emotion-Matching-Net (IMEMNet), with over 140K image-music pairs. Second, we propose cross-modal deep continuous metric learning (CDCML) to learn a shared latent embedding space which preserves the cross-modal similarity relationship in the continuous matching space. Finally, we refine the embedding space by further preserving the single-modal emotion relationship in the VA spaces of both images and music. The metric learning in the embedding space and task regression in the label space are jointly optimized for both cross-modal matching and single-modal VA prediction. The extensive experiments conducted on IMEMNet demonstrate the superiority of CDCML for emotion-based image and music matching as compared to the state-of-the-art approaches.

CCS CONCEPTS

• **Information systems** → **Sentiment analysis**; **Multimedia information systems**.

*Corresponding author.
†Equal contribution.

MM '20, October 12–16, 2020, Seattle, WA, USA

ACM ISBN 978-1-4503-7988-5/20/10…$15.00
https://doi.org/10.1145/3394171.3413776

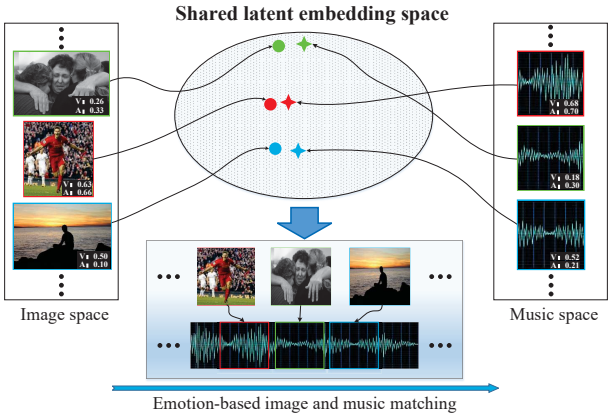


Figure 1: The basic idea of our image and music matching based on continuous emotions. Both images and music are projected into the same shared latent embedding space, which is learned by preserving both the cross-modal and single-modal emotion relationships.

KEYWORDS

Affective computing; emotion matching; valence-arousal space; deep metric learning

ACM Reference Format:
Sicheng Zhao, Yaxian Li, Xingxu Yao, Weizhi Nie, Pengfei Xu, Jufeng Yang, and Kurt Keutzer. 2020. Emotion-Based End-to-End Matching Between Image and Music in Valence-Arousal Space. In *Proceedings of the 28th ACM International Conference on Multimedia (MM '20), October 12–16, 2020, Seattle, WA, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3394171.3413776

1 INTRODUCTION

Humans are emotional animals. Famous artists remain immortal because the artworks they create such as paintings, music, and literary works can express unique insights of life and cause emotional resonance to the audience [18, 79]. The wide popularity of mobile devices and social networks enables everyone to become an

"artist". Humans habitually use images, audios, and videos together with text in social networks to express their opinions and share their emotions [82, 84]. Affective analysis of the huge volume of multimedia data can help to understand humans' behaviors and preferences and thus plays an important role in many practical applications [51, 80], such as opinion mining [23, 56, 71], business intelligence [25–27, 43, 46, 55], psychological health [6, 20, 36], and entertainment assistant [2, 54, 63, 69].

Recently, extensive efforts have been dedicated to recognizing the emotions in single-modality, such as text [19, 75], image [28, 78], speech [17], and music [68]. There are also increasing efforts on fusing information from multiple modalities [47, 52, 59, 83]. Since different modalities can provide complementary ability for emotion recognition, these multi-modal based methods usually achieve better performance. The main goal is to bridge the affective gap by extracting discriminative features and designing effective learning or fusing strategies [83].

Compared with multi-modal emotion recognition, relatively less efforts have been made to understand the emotion-centric correlation between different modalities (*e.g.* image and music studied in this paper). Such emotion-based matching is essential for various applications [57], such as affective cross-modal retrieval, emotion-based multimedia slideshow, and emotion-aware recommendation systems. The early emotion-based matching methods mainly employ a shallow pipeline [9, 34, 48, 53, 61, 86], *i.e.* extracting hand-crafted features and training matching classifiers (or training emotion classifiers for both modalities and then learning matching similarities). The differences lie in the extracted features, employed emotion representations, learned classifiers, and similarity metrics. Until very recently, some methods [57, 64] train a matching model end-to-end with specific emotion categories by concatenating the extracted visual and audio features and feeding them into a few fully-connected (FC) layers.

However, there are some limitations of existing emotion-based image and music matching methods. First, many employ limited categorical states to represent emotion. As recent psychological theories show, emotion categories in real-world are actually diverse and fine-grained [13, 74]. Therefore, such coarse-grained representation cannot well reflect the complexity and subtlety of emotions. Second, most models are trained in a multi-stage pipeline, which are impractical. Third, they do not consider how the emotional content is shared in a latent embedding space across different modalities or the learned space cannot guarantee to preserve the relationship in the label space. Finally, they do not release the datasets, which make it difficult to compare with these methods.

To address the above-mentioned problems, we propose to match image and music based on continuous valence-arousal emotions in an end-to-end manner, as shown in Figure 1. We construct Image-Music-Emotion-Matching-Net (IMEMNet), a large-scale dataset for evaluation with over 140K image-music pairs. We select DEAM [3] as the music corpora, and combine IAPS [33], NAPS [40], and EMOTIC [32] as the image corpora. To project the image and music modalities into the same shared latent embedding space, we propose cross-modal deep continuous metric learning (CDCML), which consists of three components. Cross-modal similarity metric learning enforces the distance ratios in the cross-modal matching space to be preserved in the learned embedding space. Single-modal emotion metric learning further refines the embedding space by preserving the distance ratios in the VA space of both images and music. Embedded multi-task regression learns desired regression models based on the embeddings for multi-task continuous predictions: cross-modal similarities and single-modal VA values.

In summary, the contributions of this paper are threefold:

(1) We are the first to study image and music matching based on continuous valence-arousal emotions in an end-to-end manner.

(2) We propose a novel metric learning method, CDCML, to match image and music based on emotions by learning a shared latent embedding space. The joint optimization of metric learning in the embedding space and task regression in the label space enables CDCML to simultaneously predict cross-modal matching similarities and single-modal VA values.

(3) We construct a new large-scale dataset, termed IMEMNet, for continuous emotion-based image and music matching. Extensive experimental results on IMEMNet demonstrate that the proposed CDCML method outperforms the state-of-the-art methods by a large margin for emotion-based image and music matching.

2 RELATED WORK

Emotion Representation Models. In psychology, emotion is often measured by two kinds of representation models: categorical emotion states (CES) and dimensional emotion space (DES). CES models aim to classify emotions into several discrete categories, which are easy to understand for non-professionals. The simplest CES model is sentiment polarity, *i.e.* positive and negative. More emotion categories are proposed based on psychological theories, such as Mikel's eight emotions [41] and Ekman's six emotions [16].

To more accurately model the complexity and subtlety of emotions, an increasing number of psychology studies tend to represent emotions using DES models in a 2D, 3D, or higher dimensional Cartesian space. One most popular DES model is valence-arousal-dominance (VAD) [49], where valence denotes the degree of pleasantness ranging from positive and negative, arousal shows the intensity of emotion ranging from excited to calm, and dominance represents the level of control ranging from controlled to in control. Due to the difficulty in predicting dominance, many studies represent emotions in VA space [22, 30, 85]. In this paper, we develop an end-to-end framework for cross-modal matching between image and music based on continuous emotions in VA space.

Image Emotion Recognition. The studies for image emotion recognition emerge in large numbers recently, which originates from the research in psychology to explore the relation between visual stimuli and emotion [33, 41]. In the earlier years, many types of hand-crafted representations [38, 79] are designed to bridge affective gap between low-level features and abstract emotions, such as adjective noun pairs [7, 11] and high-level concepts [4]. With the success of the convolutional neural networks (CNNs) on different multimedia tasks, current researchers mainly design CNN-based algorithms [51, 65, 67, 70, 72, 73, 81, 87]. In CES model, apart from traditional dominant emotion classification, label distribution learning [66, 76, 77] is introduced to tackle the ambiguity of image emotion by describing each category with a concrete probability. Using DES model, Kim et al. [30] developed an emotion-based network that combines low-level features, object, and background

information to predict emotion values in VA space. In [80], polarity-consistent regression loss is designed to take emotion's polarity into account for VAD prediction. Differently, our method not only penalizes the VA predictions, but also considers the feature distance in an embedding space based on the emotion similarity in VA space.

Music Emotion Recognition. Over the years, various methods have emerged to characterize and quantify the emotions associated with music. The early music emotion recognition methods mainly implement traditional machine learning algorithms with hand-crafted acoustic features as input [14, 44, 58, 60, 62], the validity and generality of which cannot be guaranteed [15]. Since these methods require careful design and data preprocessing based on extensive prior knowledge, recent emphasis has been shifted to automatically extracting features from the original data. Representative methods include CNNs [21, 39], recurrent neural networks (RNNs) especially long short-term memory (LSTM) [8, 10], and the combination of CNN and RNN [1, 15, 35]. Similar to image emotion recognition, we also preserve the music emotion similarity when learning the embedding space.

Emotion-Based Image and Music Matching. Chen et al. [9] proposed to visualize music using photos based on their emotion categories. They separately extracted hand-crafted features, learned emotion classifiers, and composited images and music based on the predicted emotions. Many methods follow this pipeline [34, 48, 53, 61, 86]. They (1) extracted more discriminative emotion features, such as low-level color [9, 34, 48, 53, 61] and mid-level principles-of-art [86] for image; (2) employed different emotion representation models, from categorical states [9, 34, 53, 61] to dimensional space [48, 86]; (3) correspondingly learned different classifiers, from Support Vector Machine [9], Naive Bayes, and Decision Tree [53] to Support Vector Regression [86]; and (4) used different composition strategies to match image and music, from emotion category comparison [9, 34, 53, 61] to Euclidean distance [48, 86].

The most relevant methods to ours are [57, 64]. Verma et al. [57] proposed to learn affective correspondence between image and music based on sentiment polarity (positive, negative, and neutral). The images and music are projected into a common representation space and a binary classification task is performed to predict the affective correspondence by a few fully-connected (FC) layers. Xing et al. [64] studied a similar task but the dataset is collected using Chinese folk images and music, which are annotated using Hevner Emotion Ring model with eight emotion categories. They also investigated the emotion similarity comparison approaches between Pearson correlation coefficient and Euclidean distance.

Differently, we propose to match image and music based on continuous emotions to better reflect the complexity and subtlety of emotions. Further, the projected latent embedding space preserves the relationship in the cross-modal similarity space and in the single-modal emotion space.

Deep Metric Learning. Deep metric learning has been widely utilized to measure the similarity or distance between different samples. As the standard loss functions, contrastive loss [12] and triplet loss [50] are milestones of deep metric learning and are widely employed in subsequent work. The contrastive loss minimizes the distance of samples from the same classes, and separates the samples of different classes away with a fixed margin. The triplet loss introduces three types of samples, named anchor, positive, and negative samples. Specifically, the loss enforces the distance between the anchor and the negative to be larger than that between the anchor and the positive. To improve the efficiency of metric learning, Oh Song et al. [42] utilized a matrix comprising pairwise distance of the mini-batch to design a loss, in which a lifted embedding structure is formed by all samples. Simultaneously, n-pair loss aims to learn the embeddings for $(n+1)$-tuple, including an anchor, a positive, and $n-2$ negative examples.

In the field of cross-modal matching or retrieval across multimedia data such as image, text, and audio, deep metric learning is broadly used to transform the features of each modality into a common embedding space [45, 88]. In [37], Liong et al. designed a unified architecture including two parallel neural networks, in which the intra-class variation is minimized and the inter-class variation is enlarged, and the difference of each sample pair from two modalities of the same class is minimized, respectively. Kang et al. [29] integrated the center loss and softmax cross-entropy loss to learn an embedding space that has a semantic meaning for both image and text for cross-modal retrieval.

As emotions in VA space are continuous values, the binary supervision that indicates whether a pair of data belong to the same class cannot describe the similarity. Inspired by log-ratio loss [31], we propose cross-modal deep continuous metric learning to measure the degree of continuous cross-modal similarity.

Table 1: Statistics of the IMEMNet dataset, where '#' denotes the corresponding number (the same below).

	Training	Validation	Testing	Total
#Songs	1,442	90	270	1,802
#Song clips	28,835	1,759	5,223	35,817
#Images	20,496	1,281	3,843	25,620
#Paires	109,525	8,795	26,115	144,435

3 THE IMEMNET DATASET

In this section, we introduce the IMEMNET dataset[1] on continuous emotion-based image and music matching, including image and music data selection and image-music matching.

3.1 Image and Music Data Selection

We combine IAPS [33], NAPS [40], and EMOTIC [32] with continuous VA labels as the image corpora. IAPS is an emotion evoking image set in psychology with 1,182 documentary-style natural color images. Each image is annotated with a 9-point VAD rating by about 100 college students. NAPS consists of 1,356 realistic, high-quality photographs rated by 204 mostly European participants in a 9-point bipolar semantic sliding scale on VA and approach-avoidance dimensions. EMOTIC is a dataset with 23,082 images containing people in non-controlled environments. The images were annotated by Amazon Mechanical Turk (AMT) workers with continuous 10-scale VAD dimensions.

We select DEAM [3] as the music corpora. DEAM consists of 1,802 excerpts and full songs annotated with VA values (from -1 to +1) both continuously (per-second) and over the whole song.

[1]The IMEMNet dataset is released at: https://github.com/linkAmy/IMEMNet.

Table 2: Comparison of our released IMEMNet dataset with others, where the values in the parentheses of the second column are the number of emotion categories or detailed emotion space, 'ED', 'AED', and 'PCC' are abbreviations for Euclidean distance, Aesthetic energy distance, and Pearson correlation coefficient, respectively.

Reference	Emotion label	#Images	#Music	Clip length	#Pairs	Matching	Released
[9]	CES (8)	368	-	5s	-	Self-defined	No
[53]	DES (VA)	3,000	1000	30s	-	ED	No
[61]	CES (3)	233	16	Unfixed	-	AED	No
[48]	DES (VA)	1,182	315	Unfixed	-	ED	No
[86]	DES (VA)	1,182	240	15s	-	ED	No
[34]	DES (VA)	57	273	20s	-	ED	No
[57]	CES (3)	85,000	3,812	60s	-	0/1	Yes
[64]	CES (8)	500	500	30s	250,000	ED & PCC	No
Ours	DES (VA)	25,620	1,802	2s	144,435	ED	Yes

Considering the stability of the annotations, each song is annotated from the 15th second. The frequency of all songs is 44100Hz. Most of the songs (1,723 in total) are 45 seconds long, with the rest varying in length, reaching a maximum of more than 600 seconds.

Since the image and music data is labeled in different scales, we normalize the VA values into [0,1] respectively based on the minimum and range. After normalization, we randomly split both image and music data into 80% for training, 5% for validation, and 15% for testing, as shown in Table 1.

3.2 Image-Music Matching

To match the images and music clips, we calculate the Euclidean distance between their VA ground truth labels, and then obtain the similarity as follows:

$$S(I_i, M_j) = \exp\left(-\frac{d\left(y^{I_i}, y^{M_j}\right)}{\sigma_n^m}\right), i = 1, \cdots, n,\ j = 1, \cdots, m, \quad (1)$$

where d stands for the Euclidean distance, y^{I_i} and y^{M_j} are the VA labels of image I_i and music clip M_j, n and m are the numbers of images and music clips, respectively. σ_n^m is set as the average Euclidean distance between all images and music clips. The degree of similarity is then set as the emotion matching label for corresponding image and music clip.

It is worth noting that all possible matching pairs is $m \times n$. For our images and music clips, the number of matching pairs will reach hundreds of millions. In order to avoid the explosion of the dataset scale, for each music clip, we select 50 images. Among them, 30 are randomly selected from the image dataset, and the remaining 20 are composed of 10 with highest matching score and 10 with the lowest. Finally, we randomly sample 10% of the pairs to constitute the IMEMNet dataset. Please note that the images and music clips of the training set, verification set, and test set do not intersect. They are constructed independently. The statistics of the IMEMNet dataset is summarized in Table 1, and the comparison of IMEMNet with existing datasets are compared in Table 2.

4 PROBLEM DEFINITION

In this paper, we study the problem of continuous emotion-based matching between images $\mathcal{I}$ and music $\mathcal{M}$, where $\mathcal{I} = \{I_i\}_{i=1}^n$ and $\mathcal{M} = \{M_i\}_{i=1}^m$. On one hand, we aim to predict the degree of similarity between images and music clips; on the other hand, we also aim to predict the concrete VA values for each modality. Given the dataset consisting of N image-music pairs $\mathcal{P} = \{(I_i, M_i)\}_{i=1}^N$ and their ground truth on the degree of similarity $\mathcal{S} = \{S(I_i, M_i)\}_{i=1}^N$, we build a branch $F_1 : \mathcal{P} \rightarrow \mathcal{S}$ to learn the similarity for the input sample pairs. Meanwhile, the single image or music clip has its own ground truth VA values. Specifically, we use $\mathcal{Y}^{\mathcal{I}} = \left\{y^{I_i}\right\}_{i=1}^n$ and $\mathcal{Y}^{\mathcal{M}} = \left\{y^{M_j}\right\}_{j=1}^m$ to represent the emotion labels of images and music clips, respectively, where $y^{I_i} = (v^{I_i}, a^{I_i})$ represent the valence and arousal of the i^{th} image and $y^{M_j} = (v^{M_j}, a^{M_j})$ represent the valence and arousal of the j^{th} music clip. Therefore, our another objective is to learn a common branch to learn the mapping $F_2 : \mathcal{I} \rightarrow \mathcal{Y}^{\mathcal{I}}$ and $\mathcal{M} \rightarrow \mathcal{Y}^{\mathcal{M}}$.

5 CROSS-MODAL DEEP CONTINUOUS METRIC LEARNING

In this section, we introduce the detailed cross-modal deep continuous metric learning (CDCML). The framework is shown in Figure 2. First, we feed images and music clips into two parallel feature extractors, which take ResNet-50 and ResNet-18 [24] as backbones, respectively. Second, we employ cross-modal similarity metric learning and single-modal emotion metric learning to learn a shared latent embedding space by optimizing various metric losses. The cross-modal similarity and single-modal emotion relationships are well preserved in the embedding space. Finally, we jointly learn a similarity predictor for image and music matching and a VA predictor for continuous emotion regression.

5.1 Cross-modal Similarity Metric Learning

In the shared feature space, we refine the feature distributions to reflect the similarity relationships and to minimize the gap between image and music modalities.

5.1.1 Cross-modal feature-ratio Loss. The popular metric learning methods usually enlarge the inter-class variation and minimize the intra-class variation by modulating the Euclidean distance between features. However, there is no specific class in VA space, where emotion label is continuous, so the widely-used metric losses cannot

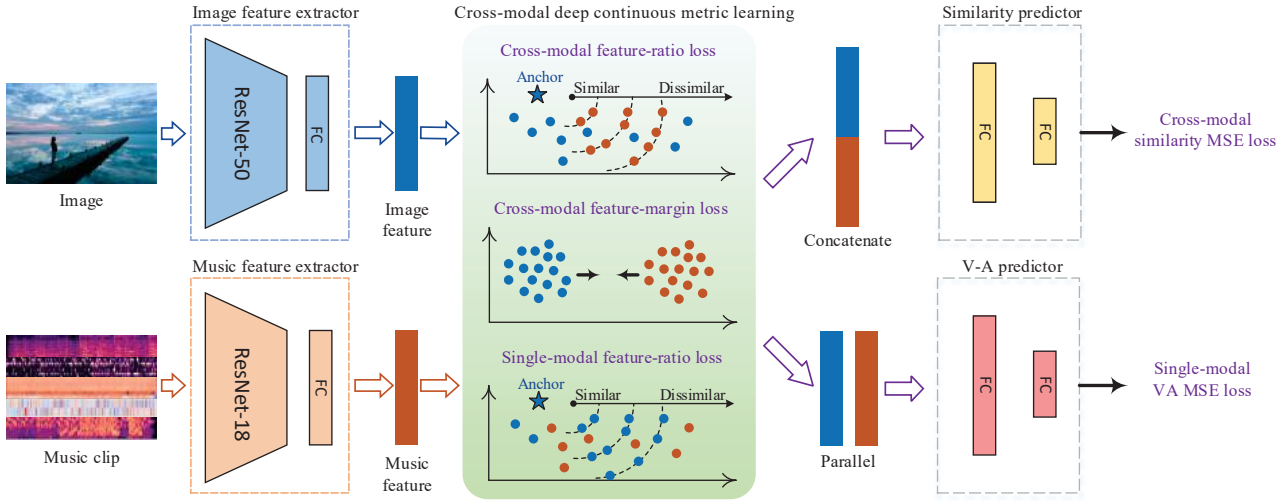


Figure 2: Framework of the proposed CDCML for continuous emotion-based image and music matching. The blue and orange circles represent image samples and music clips. 'FC' represents fully-connected layers. For simplicity, we omit the cross-modal feature-ratio loss and single-modal feature-ratio loss using music clips as anchors.

be directly applied in our tasks. Inspired by [31], we propose a cross-modal feature-ratio loss L_{CFR} to accurately optimize the distance of multi-modal embeddings based on their emotion similarity:

$$L_{CFR} = \sum_{i=1}^{N} \left\{ \log \frac{D(f^{I_i}, f^{M_i})}{D(f^{I_i}, f^{M_j})} - \log \frac{S(I_i, M_i)}{S(I_i, M_j)} \right\}^2 + \sum_{i=1}^{N} \left\{ \log \frac{D(f^{M_i}, f^{I_i})}{D(f^{M_i}, f^{I_j})} - \log \frac{S(M_i, I_i)}{S(M_i, I_j)} \right\}^2, \tag{2}$$

where $D(\cdot)$ means the squared Euclidean distance, $S(\cdot)$ denotes the similarity between image and music clip as defined in Eq. (1), and $i \neq j$. In the first item, image I_i is treated as the anchor, while the music clip M_i serves as the anchor in the second item. For the anchor of one modality (*e.g.* I_i), we randomly choose two examples from another modality (*e.g.* M_i, M_j) at one time.

5.1.2 Cross-modal feature-margin loss. To minimize the gap between the image and music spaces, we introduce cross-modal feature-margin loss to control the largest distance between representations of the two modalities. Suppose we obtain the image representation f^I and music representation f^M from their sub-networks. In order to eliminate their difference, we penalize the image-music pairs whose distances are larger than α:

$$L_{CFM} = \sum_{i=1}^{N} [||f^{I_i} - f^{M_i}||_2 - \alpha]_+, \tag{3}$$

where $[\cdot]_+ = \max(0, \cdot)$. α is a threshold to manipulate the maximum tolerable distance.

5.2 Single-modal Emotion Metric Learning

Besides the cross-modal similarity relationship, we also enforce the embedding space to preserve the continuous emotion relationships for each modality when the VA labels are available. To achieve this goal, we minimize the feature-ratio loss within each modality. Differently, the distances between features from the same modality are computed based on their Euclidean distance between VA labels. The single-modal feature-ratio losses based on image and music representations are respectively definded as:

$$L_{SFR\_I} = \sum_{i=1}^{n} \left\{ \log \frac{D(f^{I_i}, f^{I_j})}{D(f^{I_i}, f^{I_k})} - \log \frac{D(y^{I_i}, y^{I_j})}{D(y^{I_i}, y^{I_k})} \right\}^2, \tag{4}$$

$$L_{SFR\_M} = \sum_{i=1}^{m} \left\{ \log \frac{D(f^{M_i}, f^{M_j})}{D(f^{M_i}, f^{M_k})} - \log \frac{D(y^{M_i}, y^{M_j})}{D(y^{M_i}, y^{M_k})} \right\}^2, \tag{5}$$

where $D(\cdot)$ means the squared Euclidean distance and $i \neq j \neq k$. In the loss of each modality, an anchor and two neighbors take part in the loss computing. By approximating the ratio between the distances of VA labels, the learned embedding space can reflect the emotion similarity of each modality.

5.3 Embedded Multi-task Regression

After learning the shared latent embedding space, we can jointly predict the matching similities and the concrete VA values.

5.3.1 Cross-modal similarity MSE loss. The similarity predictor is used to predict the matching similarity between a pair of image and music clip. This network is composed of three fully connected layers with BatchNorm, using Relu as the activation function except for the last layer which uses Sigmoid. Taking the concatenated image-music embeddings as input, the similarity predictor aims to minimize the following mean squared error (MSE) loss:

$$L_{Sim} = \frac{1}{N} \sum_{i=1}^{N} \left(S(I_i, M_i) - \hat{S}(I_i, M_i) \right)^2, \tag{6}$$

where $\hat{S}(I_i, M_i)$ is the predicted similarity of the i^{th} image-music pair, while $S(I_i, M_i)$ is the corresponding ground truth, and N is the total amount of matching pairs.

5.3.2 Single-modal VA MSE loss. The VA predictor is used to predict the VA values of images and music clicps. This part of the network is also composed of three fully connected layers with BatchNorm, using Relu as the activation function except for the last layer which uses Sigmoid. Taking the image or music embeddings as input, the predictor minizes a similar MSE loss:

$$L_{IVA} = \frac{1}{n}\sum_{j=1}^{n}\left(y^{I_i} - \hat{y}^{I_i}\right)^2, L_{MVA} = \frac{1}{m}\sum_{j=1}^{m}\left(y^{M_j} - \hat{y}^{M_j}\right)^2, \quad (7)$$

where $\hat{y}^{I_i}$ and $\hat{y}^{M_j}$ are the predicted VA values for image I_i and music clip M_j.

5.4 CDCML Optimization

We can classify the loss functions mentioned above into two families, named similarity family $\mathcal{L}_{SF}$ and VA family $\mathcal{L}_{VAF}$. If we only have similarity labels of image-music pairs, we can use $\mathcal{L}_{SF}$, which includes L_{CFR}, L_{CFM}, and L_{Sim}, to train our framework end-to-end for matching images and music clips. If the VA labels of each modality are also available, we can simultaneously optimize both $\mathcal{L}_{SF}$ and $\mathcal{L}_{VAF}$, where $\mathcal{L}_{VAF}$ contains L_{SFR_I}, L_{SFR_M}, L_{MVA}, and L_{IVA}. Therefore, with available similarity and VA labels, our CDCML framework can be optimized by minimizing the following total loss:

$$\mathcal{L}_{CDCML} = L_{CFR}+L_{CFM}+L_{Sim}+L_{SFR\_I}+L_{SFR\_M}+L_{MVA}+L_{IVA}. \quad (8)$$

With the total loss, the embedding space and label space can be well optimized for the final prediction of multiple tasks.

6 EXPERIMENTS

In this section, we first introduce the experimental settings, including evaluation metrics, baselines, and implementation details, and then quantitatively compare the performance of the proposed cross-modal deep continuous metric learning (CDCML) method and several state-of-the-art approaches, followed by some ablation studies and visualization.

6.1 Experimental Settings

6.1.1 Evaluation Metrics. We employ mean squared error (MSE) and mean absolute error (MAE) to evaluate the effectiveness of the proposed CDCML method for image-music matching and VA prediction: $MSE = \frac{1}{t}\sum_{i=1}^{t}\left(l_i - \hat{l}_i\right)^2$, $MAE = \frac{1}{t}\sum_{i=1}^{t}\left|l_i - \hat{l}_i\right|$, where $\hat{l}_i$ represents the predicted value, l_i is the ground truth label, and t is the number of testing samples. MSE represents the sample standard deviation of the differences between predicted values and ground truth values. MAE is an arithmetic average of the absolute errors. Smaller MSE/MAE values represent better results.

6.1.2 Baselines. To compare CDCML with the state-of-the-art approaches for image and music matching, we select the following methods as baselines. **(1) SP-Net**, separately train two VA prediction models for image and music, calculate the corresponding Euclidean distance based on the predicted VA values for an image-music pair, and then obtain the matching similarity. Please note that SP-Net is trained only using VA labels. When the VA labels are unavailable, it does not work anymore. (2) L^3**-Net** [5] and (3) **ACP-Net** [57], extract features for image and music, fuse/concatenate the extracted features, and pass through several fully-connected (FC) layers to obtain the final similarity prediction. The differences lie in the input to the music feature extractors and the number of FC layers. Since L^3-Net and ACP-Net are initially designed for "general audio-visual correspondence" and "affective audio-visual correspondence" with 2-class output (*i.e.* true or false correspondence), we replace the cross-entropy loss with MSE loss. Following ACP-Net [57], we feed the learned features of both images and music clips by L^3-Net and ACP-Net into another VA predictor to compare the performance of VA prediction.

6.1.3 Implementation Details. As shown in Figure 2, our model consists of two branches: the image branch and the music branch, which are used to respectively extract visual and audio features. After metric learning, the embeddings in the shared latent space are followed by two functional sub-networks: the similarity predictor and the VA predictor.

The image branch is based on Resnet-50. We drop the original classification layer and add one additional FC layer to obtain the final 512-dimensional visual features. Each image is resized to a predefined size of [224×224×3] before passing to Resnet-50.

The Music Branch is based on Resnet-18. We also drop the classification layer and add one FC layer to extract 512-dimensional audio features. Different from images, the input of the music branch is a batch of basic music features. We first extract the [193,87]-dimensional music features, which are composed of 40 MFCCs, 12 chroma features, 7 spectral contrast features, 6 tonal centroid features, and 128 features obtained from the mel spectrogram. And then we tile the music feature to form a feature matrix in the size of [193×87×3].

The similarity predictor and VA predictor are both composed of 3 fully connected layers to respectively predict the similarity of an image-music pair and the concrete VA values of an image or music clip. Each FC layer is followed by a BatchNorm layer and an activate function layer with Relu, except for the last output layer which activation function is Sigmoid. Dropout rate is set to 0.5.

The weights of the feature extractors (*i.e.* ResNet-50 and ResNet-18) are initialized from models trained on ImageNet. The network is implemented in PyTorch and trained with SGD optimizer using a batch size of 128 with initial learning rate 1e-3. The learning rate decreases with a decay of 0.1 for every 10 epochs.

6.2 Comparison with the State-of-the-art

The comparison of the proposed CDCML method and several state-of-the-art approaches on IMEMNet is shown in Table 3. From the results, we have the following observations:

(1) SP-Net performs the worst on cross-modal similarity prediction, but obtains much better results on VA prediction than L^3-Net [5] and ACP-Net [57]. This is reasonable because SP-Net separately trains two VA prediction models for image and music. On one hand, with the VA labels as full supervision, SP-Net can learn discriminative representations for both image and music. On

Table 3: Performance of the proposed CDCML and the stare-of-art approaches on IMEMNet for continuous emotion-based image and music matching. The best results are emphasized in bold.

Method	Similarity		Image emotion				Music emotion			
	MSE	MAE	V MSE	V MAE	A MSE	A MAE	V MSE	V MAE	A MSE	A MAE
SP-Net	0.135	0.301	0.048	0.165	0.054	0.186	0.026	0.120	0.020	0.114
L^3-Net [5]	0.095	0.232	0.058	0.183	0.085	0.232	0.034	0.143	0.028	0.136
ACP-Net [57]	0.086	0.222	0.062	0.195	0.091	0.241	0.027	0.130	0.022	0.131
CDCML (Ours)	**0.067**	**0.210**	**0.044**	**0.157**	**0.050**	**0.175**	**0.024**	**0.118**	**0.015**	**0.099**

Table 4: Ablation studies of different components in CDCML on IMEMNet. 'Sim', 'VA', 'CFR', 'CFM', and 'SFR' denote the cross-modal similarity MSE loss, single-modal VA MSE loss, cross-modal feature-ratio loss, cross-modal feature-margin loss, and single-modal feature-ratio loss, respectively. '√' means the corresponding loss is utilized in the training process.

Sim	VA	CFR	CFM	SFR	Similarity		Image emotion				Music emotion			
					MSE	MAE	V MSE	V MAE	A MSE	A MAE	V MSE	V MAE	A MSE	A MAE
√					0.083	0.239	0.060	0.195	0.087	0.239	0.039	0.163	0.046	0.173
√		√			0.074	0.231	0.058	0.187	0.075	0.225	0.034	0.153	0.042	0.163
√		√	√		0.072	0.227	0.057	0.186	0.074	0.229	0.034	0.153	0.042	0.162
√	√				0.080	0.233	0.046	0.158	0.052	0.180	0.026	0.120	0.017	0.104
√	√	√	√	√	**0.067**	**0.210**	**0.044**	**0.157**	**0.050**	**0.175**	**0.024**	**0.118**	**0.015**	**0.099**

the other hand, without the similarity as supervision, it performs much worse than the methods that use similarity as supervision, *i.e.* L^3-Net, ACP-Net, and the proposed CDCML.

(2) Similar to [57], our results also show that ACP-Net [57] outperforms L^3-Net [5] on the matching and music VA prediction tasks. ACP-Net extracts various acoustic features, such as MFCC, chroma, and spectral contrast, while L^3-Net only uses log-spectrograms. Further, ACP-Net employs more FC layers to better learn the mapping between concatenated features and the similarity. However, L^3-Net performs better than ACP-Net on image VA prediction. This is because ACP-Net employs a pre-trained model to extract visual features, while the visual feature extractor in L^3-Net is trainable.

(3) CDCML obtains the best performance on both cross-modal matching and single-modal VA prediction. Specifically, compared to ACP-Net [57], CDCML achieves 22.1% and 5.4% relative performance improvements on MSE and MAE, while the relative gains over SP-Net on the valence and arousal of images and music measured by MSE are 8.3%, 7.4% and 7.7%, 25.0%, respectively. These results demonstrate the superiority of the proposed CDCML. The performance improvements benefit from the advantages of CDCML. First, it learns a shared latent embedding space which preserves the cross-similarity and single-modal emotion relationships in the label space. As a result, the embeddings are more discriminative for our task. Second, the embedded multi-task regression enables to learn a better similarity predictor and VA predictor with the joint supervision of similarity and VA labels.

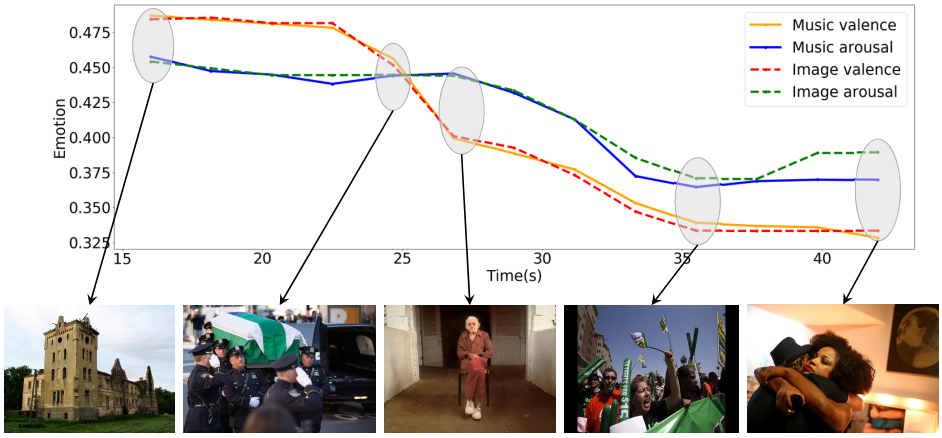
6.3 Ablation Studies

We conduct in-depth ablation studies to systematically analyze the effectiveness of different components in CDCML. The experimental results are shown in Table 4. If only cross-modal similarity labels between image and music are provided, we can train the network by optimizing cross-modal similarity MSE loss, cross-modal feature-ratio loss, and cross-modal feature-margin loss. As shown in the first part of the table, cross-modal feature-ratio loss can notably improve the performance on similarity prediction (*e.g.* 10.8% relative gains on MSE). Besides, what is pleasantly surprised is that the results of on VA prediction are also significanlty improved with the supervision of cross-modal feature-ratio loss in the shared embedding space. It demonstrates that the loss makes the feature embeddings more discriminitive not only for cross-modal matching but also for single-modal VA prediction. Note that cross-modal feature-margin loss is proposed to reduce the gap between different modalities by setting a maximum margin between features, so the performance of the output from shared FC layers is improved.

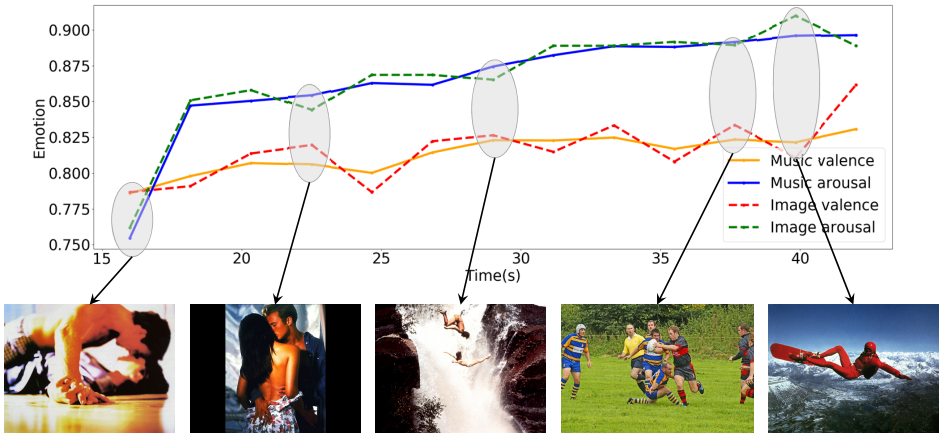
When VA labels are also provided, we can add several supervisions in both the embedding space and the VA space, as shown in the second part of Table 4. With the penalties of cross-modal similarity MSE loss and single-modal VA MSE loss on the final output, the embedded multi-task regression obtains better performance than that of using only one loss. Apart from directly using VA label in single-modal VA MSE loss, we also use single-modal feature-ratio loss to manipulate the distance between features of the same modality based on the similarity of the VA labels. It is obvious that the overall performance is further improved with the two losses that are based on VA labels, especially the results of VA prediction. The effectiveness of single-modal feature-ratio loss indicates the importance of feature distribution in the latent embedding space.

6.4 Visualization

We vividly visualize the matching results between image and music based on continuous VA emotions in Figure 3. We can observe that although the emotions of different music clips may change



(a) Low VA values



(b) High VA values

Figure 3: Visualization of emotion-based image and music matching results in the VA space (best viewed in color). For each example, the upper part shows the emotion curves of both image and music over time. In the lower part, the images with high matching similarities to corresponding music clips are shown.

dramatically, the proposed CDCML method can well match suitable images for each music clip with similar emotions.

In Figure 3 (a), the music's VA values are relatively low, indicating a negative emotion (*e.g.* sadness). It is clear that the matched images also have similar emotions. For example, the funeral and lonely elder lady both make people feel sad. In Figure 3 (b), the music's emotions are represented with high VA values, corresponding to a positive emotion (*e.g.* excitement). Meanwhile, the matched images tend to be passionate, such as the kisses and extreme sports, which can easily evoke exciting emotions. The qualitative matching results further demonstrate the effectiveness of the proposed CDCML method for matching image and music based on emotions.

7 CONCLUSION

In this paper, we aimed to study continuous emotion-based image and music matching in an end-to-end manner. To learn a shared latent embedding space, we proposed cross-modal deep continuous metric learning (CDCML) by preserving the cross-modal similarity and single-modal emotion relationships. The embedded multi-task regression can simultaneously predict the matching similarity and VA values. To evaluate the effectiveness, we constructed a large-scale dataset, termed IMEMNet. The extensive experiments on IMEMDnet demonstrate that CDCML achieves 22.1% and 5.4% relative performance improvements on MSE and MAE for matching similarity prediction as compared to the best state-of-the-art method (*i.e.* ACP-Net [57]). In future studies, we plan to model the sequential information of different music clips in a whole song using LSTM-based techniques. In addition, we will study a more practical image and music matching based on both emotions and semantics. How to deal with incremental training image-music pairs is also worth exploring.

ACKNOWLEDGMENTS

This work is supported by the National Natural Science Foundation of China (Nos. 61701273, 61876094, U1933114), Berkeley

DeepDrive, the Major Project for New Generation of AI Grant (No. 2018AAA0100403), Natural Science Foundation of Tianjin, China (Nos. 18JCYBJC15400, 18ZXZNGX00110), the Open Project Program of the National Laboratory of Pattern Recognition (NLPR), and the Fundamental Research Funds for the Central Universities.

REFERENCES

[1] Sharath Adavanne, Konstantinos Drossos, Emre Çakir, and Tuomas Virtanen. 2017. Stacked convolutional and recurrent neural networks for bird audio detection. In *European Signal Processing Conference*. 1729–1733.

[2] Hafiz Aziz Ahmad, Shinichi Koyama, and Haruo Hibino. 2012. Emotion as a Key Role in Successful Acceptance of Japanese Manga by Indonesian Readers. In *Kansei Engineering and Emotion Research International Conference*.

[3] Anna Alajanki, Yi-Hsuan Yang, and Mohammad Soleymani. 2016. Benchmarking music emotion recognition systems. *PLOS ONE* (2016), 835–838.

[4] Afsheen Rafaqat Ali, Usman Shahid, Mohsen Ali, and Jeffrey Ho. 2017. High-level concepts for affective understanding of images. In *IEEE Winter Conference on Applications of Computer Vision*. 679–687.

[5] Relja Arandjelovic and Andrew Zisserman. 2017. Look, listen and learn. In *IEEE International Conference on Computer Vision*. 609–617.

[6] Shurui Bao, Huimin Ma, and Wenyu Li. 2014. ThuPIS: A new affective image system for psychological analysis. In *IEEE International Symposium on Bioelectronics and Bioinformatics*. 1–4.

[7] Damian Borth, Tao Chen, Rongrong Ji, and Shih-Fu Chang. 2013. Sentibank: large-scale ontology and classifiers for detecting sentiment and emotions in visual content. In *ACM International Conference on Multimedia*. 459–460.

[8] Meng Cai and Jia Liu. 2016. Maxout neurons for deep convolutional and LSTM neural networks in speech recognition. *Speech Communication* 77 (2016), 53–64.

[9] Chin-Han Chen, Ming-Fang Weng, Shyh-Kang Jeng, and Yung-Yu Chuang. 2008. Emotion-based music visualization using photos. In *International Conference on Multimedia Modeling*. 358–368.

[10] Shizhe Chen, Qin Jin, Jinming Zhao, and Shuai Wang. 2017. Multimodal multi-task learning for dimensional and continuous emotion recognition. In *Annual Workshop on Audio/Visual Emotion Challenge*. 19–26.

[11] Tao Chen, Damian Borth, Trevor Darrell, and Shih-Fu Chang. 2014. Deepsentibank: Visual sentiment concept classification with deep convolutional neural networks. *arXiv:1410.8586* (2014).

[12] Sumit Chopra, Raia Hadsell, and Yann LeCun. 2005. Learning a similarity metric discriminatively, with application to face verification. In *IEEE Conference on Computer Vision and Pattern Recognition*. 539–546.

[13] Alan S Cowen and Dacher Keltner. 2017. Self-report captures 27 distinct categories of emotion bridged by continuous gradients. *Proceedings of the National Academy of Sciences* 114, 38 (2017), E7900–E7909.

[14] James J Deng, Clement HC Leung, Alfredo Milani, and Li Chen. 2015. Emotional states associated with music: Classification, prediction of changes, and consideration in recommendation. *ACM Transactions on Interactive Intelligent Systems* 5, 1 (2015), 1–36.

[15] Yizhuo Dong, Xinyu Yang, Xi Zhao, and Juan Li. 2019. Bidirectional Convolutional Recurrent Sparse Network (BCRSN): An Efficient Model for Music Emotion Recognition. *IEEE Transactions on Multimedia* 21, 12 (2019), 3150–3163.

[16] Paul Ekman. 1992. An argument for basic emotions. *Cognition & emotion* 6, 3-4 (1992), 169–200.

[17] Moataz El Ayadi, Mohamed S Kamel, and Fakhri Karray. 2011. Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition* 44, 3 (2011), 572–587.

[18] Berys Gaut. 2007. *Art, emotion and ethics*. Oxford University Press.

[19] Anastasia Giachanou and Fabio Crestani. 2016. Like it or not: A survey of twitter sentiment analysis methods. *Comput. Surveys* 49, 2 (2016), 28.

[20] Sharath Chandra Guntuku, Daniel Preotiuc-Pietro, Johannes C Eichstaedt, and Lyle H Ungar. 2019. What Twitter profile and posted images reveal about depression and anxiety. In *AAAI Conference on Artificial Intelligence*. 236–246.

[21] Yoonchang Han, Jaehun Kim, and Kyogu Lee. 2016. Deep convolutional neural networks for predominant instrument recognition in polyphonic music. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 25, 1 (2016), 208–221.

[22] Alan Hanjalic. 2006. Extracting moods from pictures and sounds: Towards truly personalized TV. *IEEE Signal Processing Magazine* 23, 2 (2006), 90–100.

[23] Syed Zohaib Hassan, Kashif Ahmad, Ala Al-Fuqaha, and Nicola Conci. 2019. Sentiment Analysis from Images of Natural Disasters. In *International Conference on Image Analysis and Processing*. 104–113.

[24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*. 770–778.

[25] Morris B Holbrook and John O'Shaughnessy. 1984. The role of emotion in advertising. *Psychology & Marketing* 1, 2 (1984), 45–64.

[26] Sameer Hosany and David Gilbert. 2010. Measuring touristsâĂŹ emotional experiences toward hedonic holiday destinations. *Journal of Travel Research* 49, 4 (2010), 513–526.

[27] Sameer Hosany and Girish Prayag. 2013. Patterns of tourists' emotional responses, satisfaction, and intention to recommend. *Journal of Business Research* 66, 6 (2013), 730–737.

[28] Dhiraj Joshi, Ritendra Datta, Elena Fedorovskaya, Quang-Tuan Luong, James Z Wang, Jia Li, and Jiebo Luo. 2011. Aesthetics and emotions in images. *IEEE Signal Processing Magazine* 28, 5 (2011), 94–115.

[29] Cuicui Kang, Shengcai Liao, Zhen Li, Zigang Cao, and Gang Xiong. 2017. Learning Deep Semantic Embeddings for Cross-Modal Retrieval. In *Asian Conference on Machine Learning*. 471–486.

[30] Hye-Rin Kim, Yeong-Seok Kim, Seon Joo Kim, and In-Kwon Lee. 2018. Building emotional machines: Recognizing image emotions through deep neural networks. *IEEE Transactions on Multimedia* 20, 11 (2018), 2980–2992.

[31] Sungyeon Kim, Minkyo Seo, Ivan Laptev, Minsu Cho, and Suha Kwak. 2019. Deep metric learning beyond binary supervision. In *IEEE Conference on Computer Vision and Pattern Recognition*. 2288–2297.

[32] Ronak Kosti, Jose M Alvarez, Adria Recasens, and Agata Lapedriza. 2017. Emotion recognition in context. In *IEEE Conference on Computer Vision and Pattern Recognition*. 1667–1675.

[33] Peter J Lang, Margaret M Bradley, and Bruce N Cuthbert. 1997. International affective picture system (IAPS): Technical manual and affective ratings. *NIMH Center for the Study of Emotion and Attention* (1997), 39–58.

[34] Taemin Lee, Hyunki Lim, Dae-Won Kim, Sunkyu Hwang, and Kyunghyun Yoon. 2016. System for matching paintings with music based on emotions. In *SIGGRAPH ASIA 2016 Technical Briefs*. 1–4.

[35] Wootaek Lim, Daeyoung Jang, and Taejin Lee. 2016. Speech emotion recognition using convolutional and recurrent neural networks. In *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*. 1–4.

[36] Huijie Lin, Jia Jia, Quan Guo, Yuanyuan Xue, Jie Huang, Lianhong Cai, and Ling Feng. 2014. Psychological stress detection from cross-media microblog data using deep sparse neural network. In *IEEE International Conference on Multimedia and Expo*. 1–6.

[37] Venice Erin Liong, Jiwen Lu, Yap-Peng Tan, and Jie Zhou. 2016. Deep coupled metric learning for cross-modal matching. *IEEE Transactions on Multimedia* 19, 6 (2016), 1234–1244.

[38] Jana Machajdik and Allan Hanbury. 2010. Affective image classification using features inspired by psychology and art theory. In *ACM International Conference on Multimedia*. 83–92.

[39] Qirong Mao, Ming Dong, Zhengwei Huang, and Yongzhao Zhan. 2014. Learning salient features for speech emotion recognition using convolutional neural networks. *IEEE transactions on multimedia* 16, 8 (2014), 2203–2213.

[40] Artur Marchewka, Łukasz Żurawski, Katarzyna Jednoróg, and Anna Grabowska. 2014. The Nencki Affective Picture System (NAPS): Introduction to a novel, standardized, wide-range, high-quality, realistic picture database. *Behavior Research Methods* 46, 2 (2014), 596–610.

[41] Joseph A Mikels, Barbara L Fredrickson, Gregory R Larkin, Casey M Lindberg, Sam J Maglio, and Patricia A Reuter-Lorenz. 2005. Emotional category data on images from the International Affective Picture System. *Behavior Research Methods* 37, 4 (2005), 626–630.

[42] Hyun Oh Song, Yu Xiang, Stefanie Jegelka, and Silvio Savarese. 2016. Deep metric learning via lifted structured feature embedding. In *IEEE Conference on Computer Vision and Pattern Recognition*. 4004–4012.

[43] Steve Pan, Jinsoo Lee, and Henry Tsai. 2014. Travel photos: Motivations, image dimensions, and affective qualities of places. *Tourism Management* 40 (2014), 59–69.

[44] Braja Gopal Patra, Dipankar Das, and Sivaji Bandyopadhyay. 2013. Unsupervised approach to Hindi music mood classification. In *Mining Intelligence and Knowledge Exploration*. 62–69.

[45] Yuxin Peng, Xin Huang, and Jinwei Qi. 2016. Cross-Media Shared Representation by Hierarchical Learning with Multiple Deep Networks.. In *International Joint Conference on Artificial Intelligence*. 3846–3853.

[46] Karolien Poels and Siegfried Dewitte. 2006. How to capture the heart? Reviewing 20 years of emotion measurement in advertising. *Journal of Advertising Research* 46, 1 (2006), 18–37.

[47] Soujanya Poria, Erik Cambria, Rajiv Bajpai, and Amir Hussain. 2017. A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion* 37 (2017), 98–125.

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

[49] Harold Schlosberg. 1954. Three dimensions of emotion. *Psychological Review* 61, 2 (1954), 81.

[50] Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In *IEEE Conference on Computer Vision and Pattern Recognition*. 815–823.

[51] Dongyu She, Jufeng Yang, Ming-Ming Cheng, Yu-Kun Lai, Paul L Rosin, and Liang Wang. 2020. WSCNet: Weakly supervised coupled networks for visual

sentiment classification and detection. *IEEE Transactions on Multimedia* 22, 5 (2020), 1358–1371.

[52] Mohammad Soleymani, David Garcia, Brendan Jou, Björn Schuller, Shih-Fu Chang, and Maja Pantic. 2017. A survey of multimodal sentiment analysis. *Image and Vision Computing* 65 (2017), 3–14.

[53] Ja-Hwung Su, Ming-Hua Hsieh, Tao Mei, and Vincent S Tseng. 2011. Photosense: Make sense of your photos with enriched harmonic music via emotion association. In *International Conference on Multimedia and Expo*. 1–6.

[54] Eduard Sioe-Hao Tan. 2008. Entertainment is emotion: The functional architecture of the entertainment experience. *Media psychology* 11, 1 (2008), 28–51.

[55] Masaki Toyama and Yuichi Yamada. 2013. Categorization of Destinations Based on TouristsâĂŹ Emotional Responses. In *TTRA International Conference*.

[56] Quoc-Tuan Truong and Hady W Lauw. 2017. Visual sentiment analysis for review images with item-oriented and user-oriented CNN. In *ACM International Conference on Multimedia*. 1274–1282.

[57] Gaurav Verma, Eeshan Gunesh Dhekane, and Tanaya Guha. 2019. Learning affective correspondence between music and image. In *IEEE International Conference on Acoustics, Speech and Signal Processing*. 3975–3979.

[58] Ju-Chiang Wang, Yi-Hsuan Yang, Hsin-Min Wang, and Shyh-Kang Jeng. 2012. The acoustic emotion Gaussians model for emotion-based music annotation and retrieval. In *ACM International Conference on Multimedia*. 89–98.

[59] Shangfei Wang and Qiang Ji. 2015. Video affective content analysis: a survey of state-of-the-art methods. *IEEE Transactions on Affective Computing* 6, 4 (2015), 410–430.

[60] Bin Wu, Erheng Zhong, Andrew Horner, and Qiang Yang. 2014. Music emotion recognition by multi-label multi-layer multi-instance multi-view learning. In *ACM International Conference on Multimedia*. 117–126.

[61] Yangyang Xiang and Mohan S Kankanhalli. 2012. A Synaesthetic Approach for Image Slideshow Generation. In *IEEE International Conference on Multimedia and Expo*. 985–990.

[62] Haishu Xianyu, Xinxing Li, Wenxiao Chen, Fanhang Meng, Jiashen Tian, Mingxing Xu, and Lianhong Cai. 2016. SVR based double-scale regression for dynamic emotion prediction in music. In *IEEE International Conference on Acoustics, Speech and Signal Processing*. 549–553.

[63] Baixi Xing, Kejun Zhang, Shouqian Sun, Lekai Zhang, Zenggui Gao, Jiaxi Wang, and Shi Chen. 2015. Emotion-driven Chinese folk music-image retrieval based on DE-SVM. *Neurocomputing* 148 (2015), 619–627.

[64] Baixi Xing, Kejun Zhang, Lekai Zhang, Xinda Wu, Jian Dou, and Shouqian Sun. 2019. Image–Music Synesthesia-Aware Learning Based on Emotional Similarity Recognition. *IEEE Access* 7 (2019), 136378–136390.

[65] Jufeng Yang, Dongyu She, Yukun Lai, and Ming-Hsuan Yang. 2018. Retrieving and classifying affective Images via deep metric learning. In *AAAI Conference on Artificial Intelligence*. 491–498.

[66] Jufeng Yang, Dongyu She, and Ming Sun. 2017. Joint Image Emotion Classification and Distribution Learning via Deep Convolutional Neural Network.. In *International Joint Conference on Artificial Intelligence*. 3266–3272.

[67] Jufeng Yang, Dongyu She, Ming Sun, Ming-Ming Cheng, Paul L Rosin, and Liang Wang. 2018. Visual sentiment prediction based on automatic discovery of affective regions. *IEEE Transactions on Multimedia* 20, 9 (2018), 2513–2525.

[68] Yi-Hsuan Yang and Homer H Chen. 2012. Machine recognition of music emotion: A review. *ACM Transactions on Intelligent Systems and Technology* 3, 3 (2012), 40.

[69] Zhengyuan Yang, Yixuan Zhang, and Jiebo Luo. 2019. Human-Centered Emotion Recognition in Animated GIFs. In *IEEE International Conference on Multimedia and Expo*. 1090–1095.

[70] Xingxu Yao, Dongyu She, Sicheng Zhao, Jie Liang, Yu-Kun Lai, and Jufeng Yang. 2019. Attention-aware Polarity Sensitive Embedding for Affective Image Retrieval. In *IEEE International Conference on Computer Vision*. 1140–1150.

[71] Jin Ye, Xiaojiang Peng, Yu Qiao, Hao Xing, Junli Li, and Rongrong Ji. 2019. Visual-Textual Sentiment Analysis in Product Reviews. In *IEEE International Conference on Image Processing*. 869–873.

[72] Quanzeng You, Jiebo Luo, Hailin Jin, and Jianchao Yang. 2015. Robust Image Sentiment Analysis Using Progressively Trained and Domain Transferred Deep Networks. In *AAAI Conference on Artificial Intelligence*. 381–388.

[73] Quanzeng You, Jiebo Luo, Hailin Jin, and Jianchao Yang. 2016. Building a Large Scale Dataset for Image Emotion Recognition: The Fine Print and The Benchmark. In *AAAI Conference on Artificial Intelligence*. 308–314.

[74] Chi Zhan, Dongyu She, Sicheng Zhao, Ming-Ming Cheng, and Jufeng Yang. 2019. Zero-shot emotion recognition via affective structural embedding. In *IEEE International Conference on Computer Vision*. 1151–1160.

[75] Lei Zhang, Shuai Wang, and Bing Liu. 2018. Deep learning for sentiment analysis: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 8, 4 (2018), e1253.

[76] Sicheng Zhao, Guiguang Ding, Yue Gao, and Jungong Han. 2017. Approximating Discrete Probability Distribution of Image Emotions by Multi-Modal Features Fusion. In *International Joint Conference on Artificial Intelligence*. 4669–4675.

[77] Sicheng Zhao, Guiguang Ding, Yue Gao, and Jungong Han. 2017. Learning Visual Emotion Distributions via Multi-Modal Features Fusion. In *ACM International Conference on Multimedia*. 369–377.

[78] Sicheng Zhao, Guiguang Ding, Qingming Huang, Tat-Seng Chua, Björn W Schuller, and Kurt Keutzer. 2018. Affective Image Content Analysis: A Comprehensive Survey. In *International Joint Conference on Artificial Intelligence*. 5534–5541.

[79] Sicheng Zhao, Yue Gao, Xiaolei Jiang, Hongxun Yao, Tat-Seng Chua, and Xiaoshuai Sun. 2014. Exploring principles-of-art features for image emotion recognition. In *ACM International Conference on Multimedia*. 47–56.

[80] Sicheng Zhao, Zizhou Jia, Hui Chen, Leida Li, Guiguang Ding, and Kurt Keutzer. 2019. PDANet: Polarity-consistent deep attention network for fine-grained visual emotion regression. In *ACM International Conference on Multimedia*. 192–201.

[81] Sicheng Zhao, Chuang Lin, Pengfei Xu, Sendong Zhao, Yuchen Guo, Ravi Krishna, Guiguang Ding, and Kurt Keutzer. 2019. Cycleemotiongan: Emotional semantic consistency preserved cyclegan for adapting image emotions. In *AAAI Conference on Artificial Intelligence*. 2620–2627.

[82] Sicheng Zhao, Yunsheng Ma, Yang Gu, Jufeng Yang, Tengfei Xing, Pengfei Xu, Runbo Hu, Hua Chai, and Kurt Keutzer. 2020. An End-to-End visual-audio attention network for emotion recognition in user-generated videos. In *AAAI Conference on Artificial Intelligence*. 303–311.

[83] Sicheng Zhao, Shangfei Wang, Mohammad Soleymani, Dhiraj Joshi, and Qiang Ji. 2019. Affective Computing for Large-scale Heterogeneous Multimedia Data: A Survey. *ACM Transactions on Multimedia Computing, Communications, and Applications* 15, 3s (2019), 93.

[84] Sicheng Zhao, Hongxun Yao, Yue Gao, Guiguang Ding, and Tat-Seng Chua. 2018. Predicting personalized image emotion perceptions in social networks. *IEEE Transactions on Affective Computing* 9, 4 (2018), 526–540.

[85] Sicheng Zhao, Hongxun Yao, Yue Gao, Rongrong Ji, and Guiguang Ding. 2017. Continuous probability distribution prediction of image emotions via multitask shared sparse regression. *IEEE Transactions on Multimedia* 19, 3 (2017), 632–645.

[86] Sicheng Zhao, Hongxun Yao, Fanglin Wang, Xiaolei Jiang, and Wei Zhang. 2014. Emotion based image musicalization. In *IEEE International Conference on Multimedia and Expo Workshops*. 1–6.

[87] Sicheng Zhao, Xin Zhao, Guiguang Ding, and Kurt Keutzer. 2018. EmotionGAN: unsupervised domain adaptation for learning discrete probability distributions of image emotions. In *ACM International Conference on Multimedia*. 1319–1327.

[88] Liangli Zhen, Peng Hu, Xu Wang, and Dezhong Peng. 2019. Deep supervised cross-modal retrieval. In *IEEE Conference on Computer Vision and Pattern Recognition*. 10394–10403.