EURASIP Journal on Audio, Speech, and Music Processing

METHODOLOGY

Open Access

YuYin: a multi-task learning model of multi-modal e-commerce background music recommendation

Le Ma[1], Xinda Wu[1], Ruiyuan Tang[1], Chongjun Zhong[1] and Kejun Zhang[1,2*]

Abstract

Appropriate background music in e-commerce advertisements can help stimulate consumption and build product image. However, many factors like emotion and product category should be taken into account, which makes manually selecting music time-consuming and require professional knowledge and it becomes crucial to automatically recommend music for video. For there is no e-commerce advertisements dataset, we first establish a large-scale e-commerce advertisements dataset Commercial-98K, which covers major e-commerce categories. Then, we proposed a video-music retrieval model YuYin to learn the correlation between video and music. We introduce a weighted fusion module (WFM) to fuse emotion features and audio features from music to get a more fine-grained music representation. Considering the similarity of music in the same product category, YuYin is trained by multi-task learning to explore the correlation between video and music by cross-matching video, music, and tag as well as a category prediction task. We conduct extensive experiments to prove YuYin achieves a remarkable improvement in video-music retrieval on Commercial-98K.

Keywords Cross-modal retrieval, Multi-modal, Music recommendation

1 Introduction

Background music (BGM) plays a vital role in advertisements, which can help build brand image and stimulate consumption [1–3]. Many studies from psychology and brain science have been carried out on the effect factors of BGM. By observing the brain, these studies have proven that BGM is associated with faster response times and greater activations of frontoparietal areas during happy music, whereas sad music is associated with slower responses and greater occipital recruitment. When the emotion of BGM is in path with the advertisement, it can help catch the attention of customers [4] and makes the advertisement more memorable [5, 6]. However, with the expanding demand for e-commerce advertisements, manually selecting music one by one and clipping the music not only requires professional knowledge but is also time-consuming from the ever-growing music pool, which makes it a crucial task for automatically selecting suitable BGM.

Recommending appropriate music for a video can be considered a cross-modal retrieval task, aiming to search relevant data in different formats [7]. Previous studies have mainly focused on retrieval between visual and textual modalities, such as retrieving images or videos corresponding to a given textual description [8–10] or generating textual descriptions for a given image or video [11–13].

Among existing video-audio retrieval research, some studies focus on sound events localization [14, 15], which aims to localize the object in the video that produces the sound. Other studies concentrate on face-speech retrieval, which seeks the corresponding person for a

*Correspondence:
Kejun Zhang
zhangkejun@zju.edu.cn
[1] Zhejiang University, Hangzhou, China
[2] Innovation Center of Yangtze River Delta, Shanghai, China

given voice [16–18]. However, there are several challenges in video-music retrieval. First, there are limited public datasets for video-music retrieval. The datasets used in existing studies are mainly music videos from YouTube. Second, there is no explicit correlation between the video and music.

Considering that music primarily depends on the video's emotion, many studies have relied on emotion tags [19, 20], which are time-consuming to annotate and may introduce subjective bias. Then, some studies use the content-based model to directly learn the correlation between video and music through deep neural networks (DNN) by calculating the Euclidean distance or cosine similarity between video and music features [21–23]. However, the music features are coarse in these studies, and emotion features may be ignored.

In this paper, to stress these challenges, we first establish a large-scale multi-modal dataset Commercial-98K from Alibaba, covering major product categories. Moreover, we propose a content-based video-music retrieval model YuYin. Instead of emotion tags, we extract emotion-related features from music, which avoid the subjective bias in annotating. Also, we introduce a weighted fusion module (WFM) to fuse the emotion features and the audio features for more fine-grained music representation, which can dynamically weigh different features, thus reducing information redundancy and enhancing the robustness of the model.

For the background music also relevant to the product category [24–26], we apply multi-task learning to train YuYin, including the cross-matching task and the category prediction task. Specifically, for the cross-matching task, the text features of the category tag are used to help align video and music. For the category prediction task, a weight-shared classifier is used to predict the category of videos, music, and text.

The main contributions of this paper can be summarized as follows:

- We establish a large-scale dataset, Commercial-98K, containing large-scale advertisements from Alibaba and covering major product categories.
- We propose a novel video-music retrieval model, YuYin, trained by multi-task learning, with categories included as labels to be predicted and as a supportive modality to align related video and music.
- We introduce a weighted fusion module to fuse emotion features and audio features of music for fine-grained music representations, learning to dynamically balance different features through training.

The rest part of the paper is organized as follows. We discuss related work in the field of multi-modal datasets and video-music retrieval in Section 2. Then, we explain the process of building our dataset Commercial-98K, including the data sources and the details of the data processing, in Section 3. Our proposed model YuYin is depicted in detail in Section 4. Then, we introduce the experiment setup and analyze the results in Section 5. Finally, we present the conclusion in Section 6.

2 Related Work

2.1 Multi-modal Dataset

Compared with single-modal datasets, multi-modal datasets contain more than one data form and are proven to have more advantages. DEAP [27] includes self-assessment scores, audio, videos, facial expressions, and physiological data for analyzing human emotional states, which shows the improvement in the effectiveness of human emotion analysis. VQA [28] containing 25,000 images, 7600 questions, and 100,000 answers. Many studies on VQA achieve better results in the tasks of free-form and open-visual question answering (Table 1).

In the field of cross-modal retrieval, researchers have built multi-modal datasets with different scales, modalities, and sources for specific tasks (Table 1). YouTube-8M [35] is released by Google, which is one of the largest multi-modal datasets. YouTube-8M contains 8,000,000 videos and text annotations from Youtube, which are divided into 4800 categories, and each video contains an average of 1.8 tags. Based on Youtube-8M, there are many subsets like HIMV-200k [21], which contains 205,000 music video-audio pairs. UGV [30] is a video dataset with emotion tags and is used for music recommendation. HoK400 and CFM400 [29] are two game video datasets

Table 1 Multi-modal datasets created and adopted in existing studies

Dataset	Scale	Modal			Content
		V	M	T	
CFM400 [29]	401	✓	✓		Game videos (Cross fire)
HoK400 [29]	427	✓	✓		Game videos (Honor of king)
UGV [30]	1265	✓	✓	✓	User generated videos
YouCook2 [31]	2000	✓		✓	Cooking videos on Youtube
EmoMV [32]	5986	✓	✓	✓	Music videos with emotion label
MSR-VTT [33]	10,000	✓		✓	Online videos with caption
TT-150K [34]	150,000	✓	✓	✓	Microvideos on Tiktok
HIMV-200K [21]	205,000	✓	✓		Music videos on YouTube
Youtbe-8M [35]	8,000,000	✓	✓	✓	Videos on YouTube
Commercial-98K	98,071	✓	✓	✓	E-commerce ads

The modalities corresponding to abbreviations in the table are as follows: *V*, video; *M*, music; *T*, text

established from the short video platform and add voice-over in the dataset besides video and background music. TT-150K [34] is a large-scale dataset established from TikTok for background music recommendation, which contains 150,000 user-generated short videos corresponding to 3000 pieces of background music.

However, the datasets for video-music retrieval are mainly music videos that are consciously made for the specific music, which makes the diversified demand difficult to achieve and can hardly fit the e-commerce scenario and there are still problems of uneven length and quality of audio and video. Therefore, we establish a multi-modal dataset Commercial-98K, containing videos, music, and tags from the top store in the largest Chinese e-commerce platform Alibaba.

2.2 Video-music retrieval

Cross-modal retrieval (CMR) aims to retrieve data between multiple modalities [7]. While there have been many studies on visual-text retrieval along with public datasets, such as Flicker [36], HowTo100M [37], and YouCook2 [31]. However, limited studies have focused on video-music retrieval (VMR).

Compared to visual-text retrieval, VMR is more challenging because both video and music contain rich information, which makes the "modality gap" rather huge. To bridge the modality gap between video and music, in paper [19], they notice that videos have a strong connection with the emotion of the BGM and perform music retrieval by calculating the textual similarity of their emotion tags. However, these emotion tags are mainly annotated through crowd-sourcing, while the quality of labels can hardly be guaranteed and may introduce subjective bias.

Then, Some studies use the content-based model to directly learn the correlation between video and music. The shallow model like CCA [38] has been used for correlation analysis between different modalities, in which a linear projection is learned to map different modalities in the same space and maximize their correlation. Based on CCA, DCCA [39] is proposed, which extends CCA with non-linear projections by DNN. In paper [40], a CCA-based model is used to match two modalities by maximizing the correlation between image and music features.

Recently, deep metric learning has been widely applied in cross-modal retrieval [41–45], which aims to learn a function to calculate the metric, e.g., euclidean distance, inner product, or cosine similarity between different modalities.

Normally, features of different modalities are extracted first and projected into a common space before calculating the metric [21, 23, 46, 47]. For video, the features are usually extracted by a pre-trained convolutional neural network (CNN) [22, 48]. However, the music representations are more complex. In [21], handcrafted features like Mel-spectrogram and MFCC are designed to represent music. Some other studies use pre-trained networks like VGGish [49] to extract audio features [23, 29, 34]. To obtain the emotion of music, In [22], they improve the feature extractor by pre-training on the emotion classification task. CMVAE [34] extract emotion features by OpenSmile [50] as well as audio features by VGGish, then they apply concatenation and principal components analysis (PCA) to obtain the final music features.

After feature extraction, this study employs pair-wise metric learning to explore the correlation between different modalities [37, 51, 52]. Specifically, positive and negative pairs are constructed for video and music, followed by pair-wise loss functions, such as noise-contrastive estimation (NCE) [53] and Triplet-loss [54], to minimize the distance between positive pairs while pushing negative pairs away. Additionally, CEVAR [23] constrains the distance between the video and music from the same clip by cosine similarity loss. In the work of Liu et al. [30], videos and music are first categorized into positive and negative pairs based on whether they have the same emotion label, and a DNN is utilized to align the two modalities. Moreover, CBVMR [21] utilizes intra- and inter-constraints to obtain more fine-grained information between video-music pairs. Generative methods are also applied in the field of CMR. Zhou et al. [55] propose an end-to-end model to generate sound for videos. Additionally, CMVAE [34] is based on a variational auto-encoder (VAE) to cross-generate music and short videos, besides learning the correlation between different modalities in the latent space. However, most studies only consider videos and music, while the text also contains valuable information for video-music retrieval. Although CMVAE [34] fuses text features with video features, the performance is likely to degrade when the text is absent.

Therefore, we propose a content-based model YuYin, which learns the correlation between video and music by multi-task learning. For better music representation, YuYin extracts and fuses the emotion features and the audio features from music. Besides, text features are also included in multi-task learning which helps to gather the video and music in the same category in the common space. The text is only used as a supportive modality in the training phase and does not involve in the video-music retrieval, thus the missing text does not have any impact on the model.

3 Dataset

The main aim of the Commercial-98K Dataset is to bridge the gap that there is no dataset that associates advertisements with background music and facilitates research regarding the discovery of matching patterns between music and advertisements. In this section, we in detail discuss the steps we took to collect advertisements and background music from the e-commerce platform as well as the data preprocess methods. Then, we depict the statistics of the collected advertisements and music.

3.1 Data collection

We collect advertisements from Taobao, one of the largest e-commerce platforms in China belonging to the world-famous company Alibaba, where customers can buy and sell numerous products. Compared with other e-commerce platforms, Taobao is a customer-to-customer e-commerce platform, where both enterprises and individuals can open online stores to sell their products. The stores on Taobao upload their products with basic information as well as some images or videos and categorize them into different sections, which produces a vast amount of advertisements in different categories. However, the advertisement quality on Taobao differs vastly, for enterprises upload advertisements made by professionals while many individuals just casually make videos introducing their products. Also, many advertisements only contain voice-over, which is not helpful and may introduce noisy data in video-music correlation learning. To ensure the quality of our data, we primarily collect advertisements from brand stores. These advertisements are designed, shot, and edited by professionals with careful attention paid to the tight correlation between the video and the background music. We ultimately gather 11,500 advertisements from 15 categories on Taobao including food, children's clothing, tablets, wedding dresses, women's t-shirts, men's t-shirts, men's suits, video games, women's suits, baby products, daily necessities, sports, down jacket, cosmetics, mobile phones.

3.2 Data preprocess

With the collected 115,000 e-commerce advertisements, as shown in Fig. 1, we separate the audio and visual content of the video by moviepy to find some audios are primarily voice-over or muted. To filter these voice-over or mute audios, we use a pre-trained time-domain convolution network [56] to calculate the onset and duration proportion of music in audios and exclude data where music accounts for less than 50% of the total time. Finally, as in Fig. 2a, we retain 98,071 advertisements in 15 categories and find the count of advertisements categories varies significantly. Hence, we further manually merged the advertisements from similar categories, e.g., tablets, mobile phones, and video games are merged as electronic products.

Then we obtain four categories, namely electronic products (video games, mobile phones, tablets), baby products (baby care, children's clothing), daily necessities (daily necessities, food, cosmetics), and clothing (sports, wedding dresses, women's t-shirts, down jacket, women's suits, men's suits, men's t-shirt).

3.3 Data statistics

As depicted in Fig. 2b, the 98,071 advertisements consist of 4 categories, namely 43,841 on clothing, 25,860 on baby products, 26,824 on daily necessities, and 1546 on electronic products. Commercial-98K is still unbalanced, for the count of data in electronic products is notably lower. The reason may be that, compared with other categories, the number of brand stores on Taobao for electronic products is mainly famous brands at home and abroad, of which the number is limited. Due to copyright restrictions, we can not propose the raw data but the processed Commercial-98K can be downloaded on https://github.com/Venatoral/Commercial-98K.

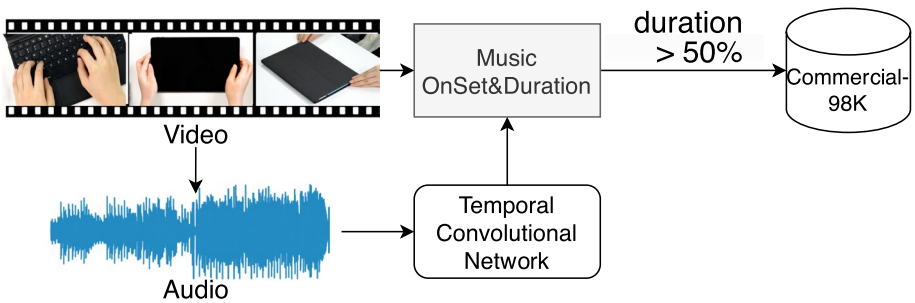
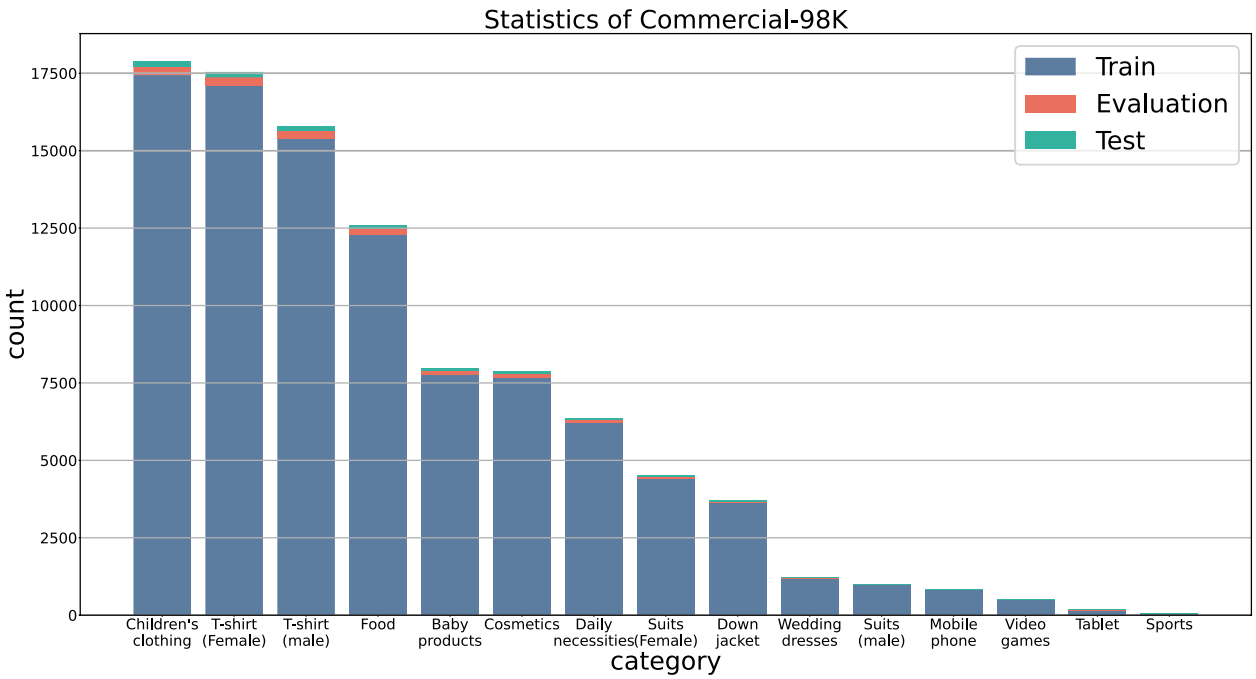
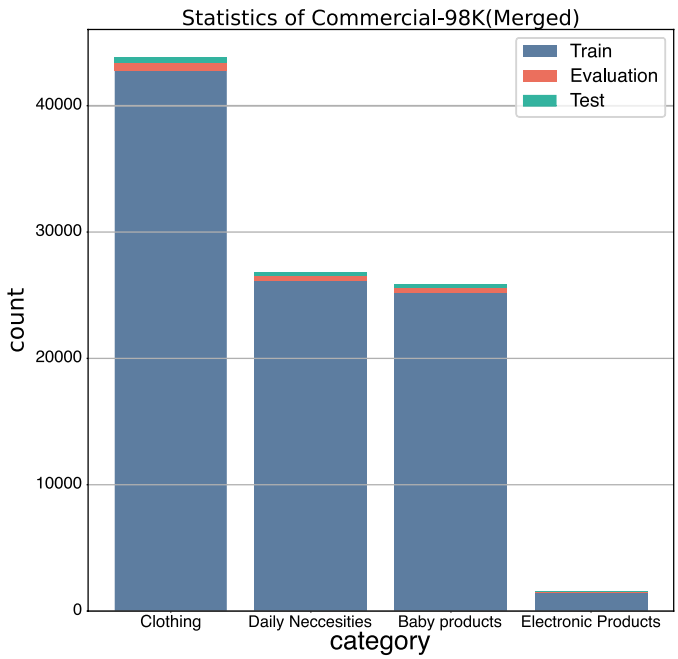


Fig. 1 The data process pipeline of Commercial-98K



(a) Original Commercial-98K.



(b) Merged Commercial-98K.

Fig. 2 The data distribution of Commercial-98K before (left panel) and after (right panel) merging the categories

4 Proposed method

4.1 Problem definition

Let M stand for a collection of music and V for a collection of advertisements. The video features v in V are extracted from the frame image sequence. Thus, it is possible to define a function $f : M \times V \rightarrow S$, where S stands for the similarities matrix and each $s_{ij} \in S$ denotes the similarities between the ith piece of music ($m_i \in M$) and

the jth advertisement ($v_j \in V$). Given a new advertisement v and the function f, the candidate music set $C(m)$ from M can be selected by computing and scoring the similarities between v and each music clip $m \in M$.

4.2 Overall framework

As the framework of our proposed video-music retrieval model YuYin shown in Fig. 3, we extract emotion features and audio features from the music separately, while the video features and text features are extracted from the sampled image sequences and tags. Through the WFM, the emotion features and audio features are fused to be the music features. Then, different features are projected into the common space. Thus, multi-task learning is applied to compute the cross-matching loss as well as prediction loss to learn the correlation between advertisements and music. Eventually, by computing the cosine similarities between video and music in the common space and ranking, the candidate music for the given advertisements can be selected.

4.3 Feature extraction

We use multiple pre-trained networks as feature extractors for different modalities. Furthermore, for stability, all feature extractors are frozen during training.

For music, we extract the frequency domain features from the music clip by torchaudio [57] as the input of pre-trained AST [58], to obtain the audio features. Besides, we apply OpenSmile [50] to extract emotion features from music clips with its emobase feature set.

For video, we sample frames from the videos at a certain rate, then the sampled frames are fed into the pre-trained inception [48] to get the frame-level features. Finally, inspired by the work [35], we use temporal global average pooling to obtain the video-level features.

For text, since the advertisements are from Chinese e-commerce platforms, we use Bert-wwm [59], which is pre-trained on the Chinese wiki, to extract the text features from the tag of the advertisement.

4.4 Weighted fusion module

As shown in Fig. 4, we introduce a more flexible fusion method called the weighted fusion module to get the music features m from the audio features a and emotion features e. The dynamic weights ranging from 0 to 1 are learned for concatenated features through the linear and sigmoid layers. Eventually, to reduce the dimension of the weighted features, a linear layer is applied to output the music feature.

4.5 Multi-task learning

YuYin is trained through multi-task learning. First, YuYin uses pair-wise metric learning to learn the direct relationship between different modalities. We set videos and music clips from the same advertisement as the positive pairs and others as the negative pairs. With the positive and negative video-music pairs constructed, NCE loss is applied to learn the correlation between the video-music pairs, as described in Eq. (1), where x and y stands for two different modalities, $P(x)$ means the positive data of x, B is the batch size, and τ

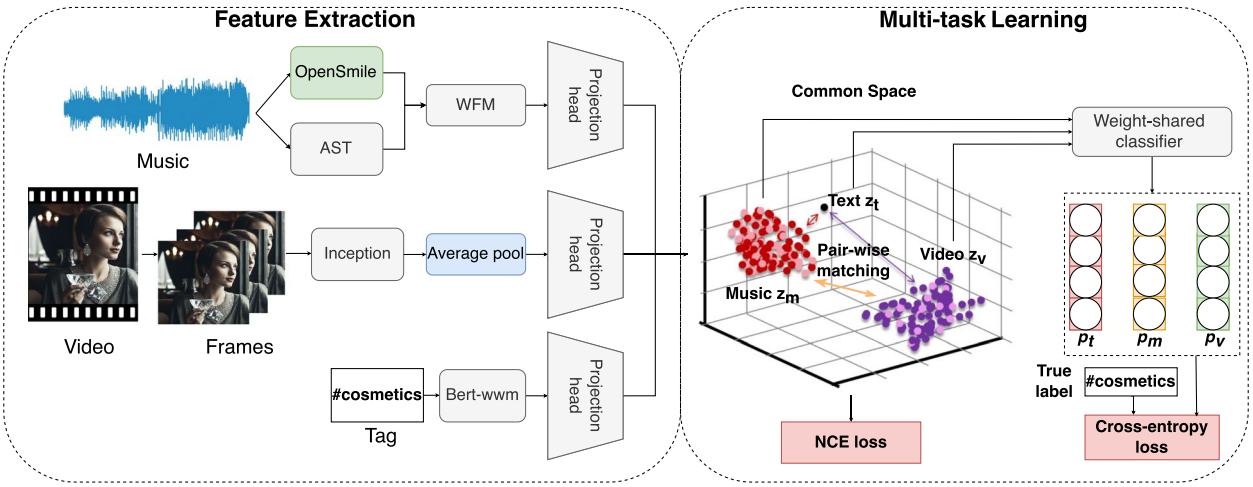


Fig. 3 The framework of our proposed YuYin for background music recommendation of e-commerce advertisements. In detail, a WFM fuses emotion features and audio features as music features. Then the extracted features are projected in the common space for multi-task learning. The video z_v, music z_m, and text projections z_t in the common space are pair-wise cross-matched to compute NCE loss and pass through a weight-shared classifier to get the prediction probabilities p_v, p_m, and p_t, which will be further used to compute cross-entropy loss with the true label as the prediction loss

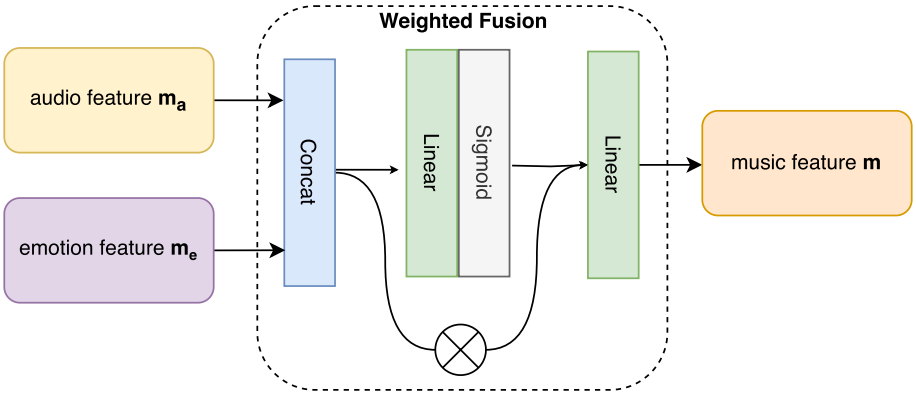


Fig. 4 The weighted fusion module (WFM) in YuYin, which learns to apply dynamic weights for audio and emotion features and output the music feature

is a hyper-parameter. Besides, the text features are included to align video and music. Specifically, the text features of the tag are extracted and projected into the common space to match the corresponding videos and music by Eq. (1).

Equation (2) illustrates how video projections z_v, music projections z_m, and text projections z_t yield the cross-matching loss L_{cm}, where β is a hyper-parameter used to regulate the video-music matching loss. Through optimization, the distance between the positive video-music pairs in the common space steadily decreases, while the distance between the negative pairs keeps growing.

$$NCE(x,y) = -\log \frac{\sum_{y \in P(x)} e^{x^T y/\tau}}{\sum_{i=1}^{B} e^{x^T y/\tau}} \tag{1}$$

$$\begin{aligned} L_{cm} = \beta * NCE(z_v, z_m) + NCE(z_v, z_t) \\ + NCE(z_m, z_t) \end{aligned} \tag{2}$$

Additionally, we provide a category prediction task to aid YuYin in learning the relationship between video and music in the same product category. The prediction loss L_{pre} is computed as Eq. (4), where CE is the cross-entropy loss and y is the ground-truth label. Specifically, a weight-shared classifier predicts the label of various modalities in the common space separately. By optimizing L_{pre}, the correlation between the videos and music with the same label is better exploited, reducing the distance between positive video-music pairs.

$$CE(x,y) = -\sum y \log(x) \tag{3}$$

$$L_{pre} = CE(p_v, y) + CE(p_m, y) + CE(p_t, y) \tag{4}$$

Eventually, the loss L consists of cross-matching loss L_{cm} and prediction loss L_{pre}. The α is a hyper-parameter to control the impact of prediction loss.

$$L = L_{cm} + \alpha * L_{predict} \tag{5}$$

5 Experiment

5.1 Experiment setup

For video-music retrieval methods can only retrieve music from the music pool without editing, which may cause misjudgment in subjective evaluation because listeners can hardly know how the music will be used as BGM of the given video, we only conduct objective experiments on Commercial-98K. We conduct experiments on Commercial-98K, with 95,607 data serving as the training set, 1464 as the testing set, and the remaining 1000 as the evaluation set. In addition, each set includes all of the dataset's categories.

YuYin is implemented in Pytorch with an embedding dimension of 1024 and the common space projection using a MLP with two layers of dimensions {512, 256} and activation function ReLu. α in Eq. (5) is set to 0.1, while β in Eq. (2) is set to 3.0. YuYin is trained on RTX3090 for 30 epochs using the Adam optimizer, with a batch size of 1024 and a learning rate of 0.0001. Following each epoch, the model is evaluated on the evaluation set to determine the evaluation loss, which is observed to prevent overfitting.

5.2 Evaluation metrics

As the standard cross-modal retrieval metric, *Recall@K* is used to validate the performance of YuYin on the video-music retrieval task [60]. As shown in Eq. (6), *Recall@K* denotes the top K retrievals obtained from the similarity list retrieved by the model, sorted in descending order $S[:K]$ as a ratio of the number of hits to the number of queries N_{query}.

$$Recall@K = \sum_{S[:K]} \frac{hits}{N_{query}} \tag{6}$$

5.3 Performance comparison

In this study, we compare YuYin with several video-music retrieval methods below:

- CCA [38]: CCA uses a linear projection and maximizes the correlation between the latent variables of video and music during training.
- DCCA [61]: DCCA learns the projection for each modality and maximizes their correlation through deep learning.
- CEVAR [23]: CEVAR uses two sets of fully-connected networks (FC) to extract video features and audio features in Youtube-8M to calculate cosine loss and predict the label of video as the prediction loss. We maintained its strategy to use the tags in Commercial-98K for its prediction loss.
- CBVMR [21]: CBVMR is a content-based video-music retrieval model, which introduces intra- and inter- modality constraints on the audio features and the video features.
- CMVAE [34]: CMVAE is based on the VAE architecture, which fuses the video features and text features through a Product-of-Expert (PoE) module and projects the fused video and music features into a latent space to compute reconstruction loss and cross-matching loss for training. For comparison, we retrain CMVAE on Commercial-98K and use the tags in Commercial-98K as the text features for fusion.
- MRCMV [29]: MRCMV fuses voice-over with video features through a multi-head attention module and uses two separate self-attention modules for the video and music features. In our comparison, for there is no voice-over in Commercial-98K, we replace the voice-over features with our text features.
- Random: randomly recommend music for the given video.

The results are shown in Table 2, which indicates that YuYin outperforms other methods on Commercial-98K. The performance of CCA indicates that the correlation between the videos and music is difficult to learn with a linear projection. CBVMR has inter- and intra- modality constraints, but the hand-crafted audio features can represent limited information. Although CEVAR introduces labels for prediction besides computing the cosine loss between the videos and music, fine-grained information of video and music can hardly be exploited through two sets of fully connected networks. The performance of CMVAE on video-music retrieval and music-video retrieval is equally well, which may be attributed to the cross-matching and reconstruction loss used in the training stage to help it catch more correlation between the videos and music. However, we also find that even though we replace the voice-over in MRCMV with our text features, it also shows considerable results. The reason may be that the multi-head attention module and the self-attention module in MRCMV refine the video features and music features, which makes MRCMV apply more task-related information.

5.4 Ablation study

In this section, we explore the specific impact effects of each component in YuYin. First, we investigate the effect of text and emotion in YuYin by eliminating and retraining. Then, we replace WFM in YuYin with traditional fusion approaches. Moreover, we investigate the effect of multi-task learning on the video and music features in the common space through feature visualization.

5.4.1 Effect of emotion and text

We verify the effect of each modality by eliminating the text features (YuYin w/o T) and the emotion features

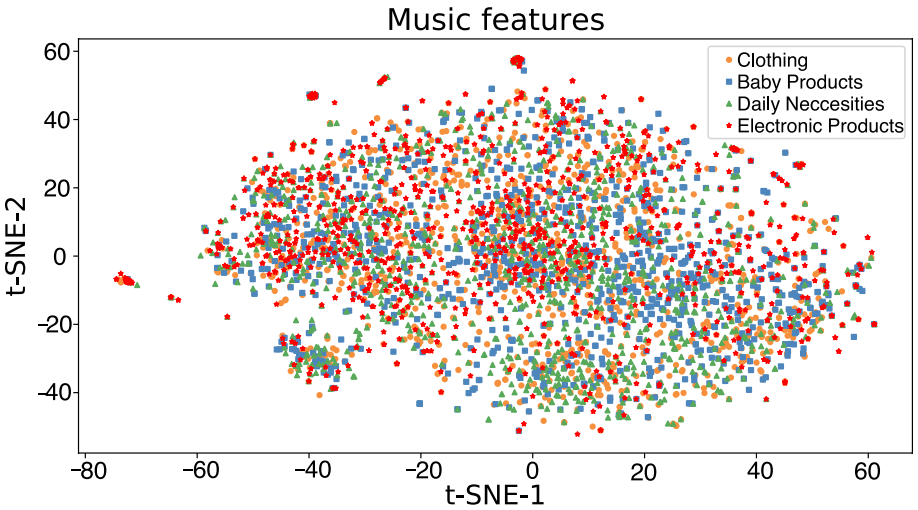
Table 2 The results of YuYin and other compared methods on Commercial-98K dataset

Methods	Video → music				Music → video			
	R@10	R@15	R@20	R@25	R@10	R@15	R@20	R@25
Random	0.017	0.023	0.026	0.030	0.013	0.021	0.026	0.033
CCA [38]	0.255	0.283	0.299	0.324	0.236	0.260	0.279	0.298
DCCA [61]	0.279	0.328	0.364	0.393	0.265	0.300	0.329	0.361
CEVAR [23]	0.334	0.362	0.378	0.389	0.335	0.359	0.372	0.392
CBVMR [21]	0.343	0.378	0.405	0.423	0.297	0.329	0.355	0.387
CMVAE [34]	0.343	0.379	0.408	0.442	0.339	0.377	0.414	0.435
MRCMV [29]	0.359	0.404	0.431	0.455	0.333	0.375	0.432	0.435
YuYin	**0.403**	**0.439**	**0.471**	**0.501**	**0.376**	**0.423**	**0.456**	**0.478**

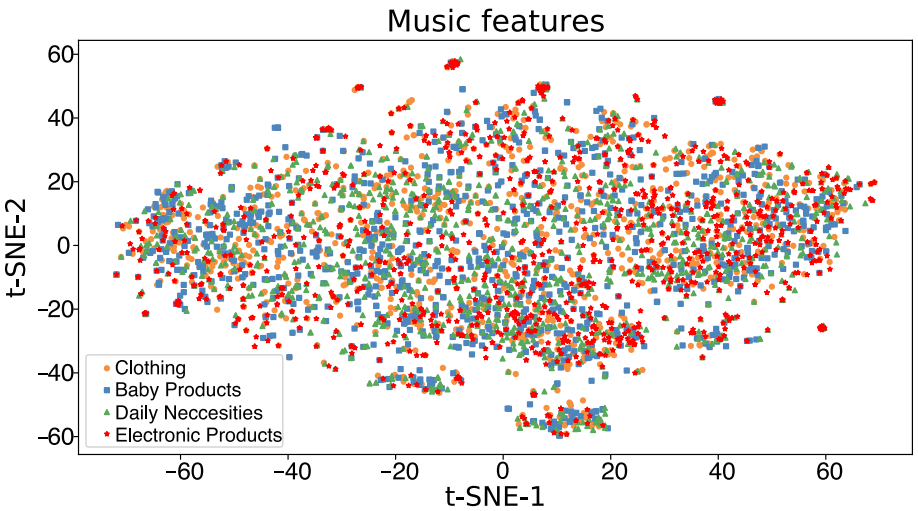
(YuYin w/o E), and YuYin pure only uses audio and video features. When eliminating emotion features, the related WFM is also removed. Furthermore, when eliminating text features, we also remove its related cross-matching loss while keeping the prediction loss. From the results in Table 3, removing either emotion or text decreases the performance of YuYin, among which the emotion features have the greatest impact on YuYin. The result may

Table 3 The results of YuYin that eliminating modalities on Commercial-98K

Methods	Video → music				Music → video			
	R@10	R@15	R@20	R@25	R@10	R@15	R@20	R@25
YuYin pure	0.370	0.400	0.424	0.443	0.345	0.378	0.400	0.417
YuYin w/o E	0.374	0.403	0.430	0.453	0.345	0.377	0.404	0.424
YuYin w/o T	0.368	0.406	0.445	0.477	0.358	0.390	0.405	0.428
YuYin	**0.403**	**0.439**	**0.471**	**0.501**	**0.376**	**0.423**	**0.456**	**0.478**



(a) Visualization of the music features of YuYin



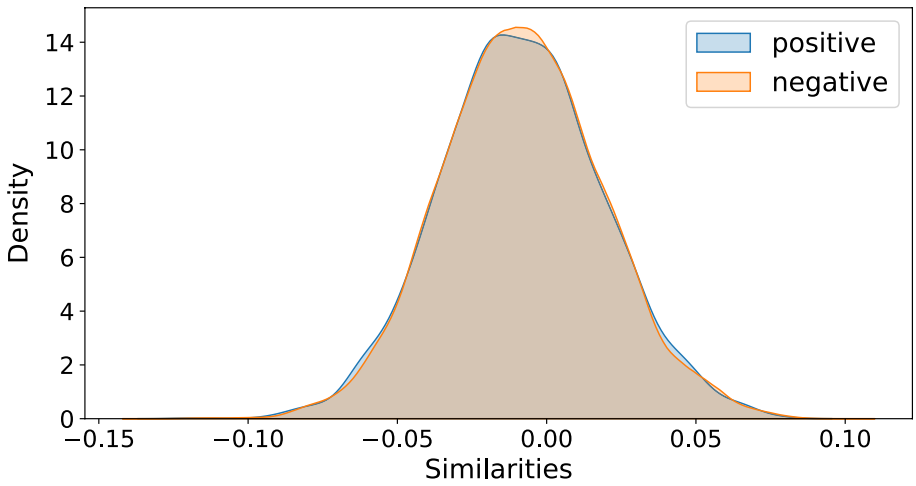
(b) Visualization of the music features of YuYin w/o E

Fig. 5 The visualization of music features from YuYin (the left panel) and YuYin w/o E (the right panel), of which the dimension is reduced to 2 by t-distributed stochastic neighbor embedding (t-SNE)

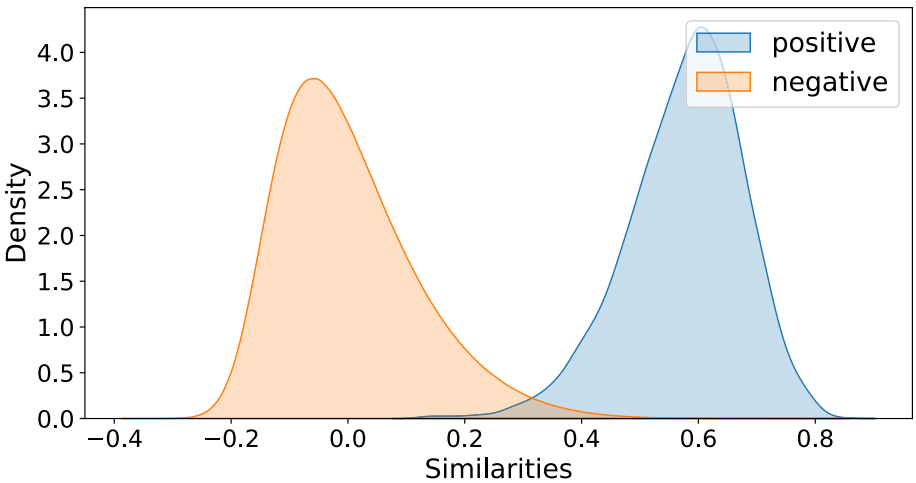
be attributed to that the emotion features extracted by OpenSmile have more intuitive meaning than the audio features extracted by AST. However, we also find the text features have less impact on the performance of YuYin. To explore the reason why the text features can hardly improve the performance when eliminating the emotion features, we randomly extract features from 1000 music in each category in Commercial-98K and reduce the dimension of music features to 2 by t-distributed stochastic neighbor embedding (t-SNE) for visualization. As

Table 4 The results of YuYin with different fusion approaches on Commercial-98K

Methods	video → music				music → video			
	R@10	R@15	R@20	R@25	R@10	R@15	R@20	R@25
YuYin (Add)	0.376	0.402	0.423	0.445	0.348	0.381	0.402	0.425
YuYin (Concat)	0.374	0.406	0.432	0.451	0.344	0.378	0.410	0.428
YuYin (WFM)	**0.403**	**0.439**	**0.471**	**0.501**	**0.376**	**0.423**	**0.456**	**0.478**



(a) KDE of the video and music features.



(b) KDE of the video and music projections.

Fig. 6 The KDE analysis of the similarity between the positive and negative samples

shown in Fig. 5, we can observe that the music features become more discriminative with the emotion features, while the music features in YuYin w/o E are sparse. Then, we analyze that the reason may be that the text features only act as a supportive role in aligning video and music in the training phase, and the sparse music features make it hard for text features to align other modalities, resulting in the performance of YuYin pure comparable to that of YuYin w/o E. In addition, the results further prove the effect of emotion features, the reason why audio features are less different compared with emotion features may be attributed to that AST is frozen during the training phase, while OpenSmile extracts emotion features from its fixed rules.

5.4.2 Effect of WFM

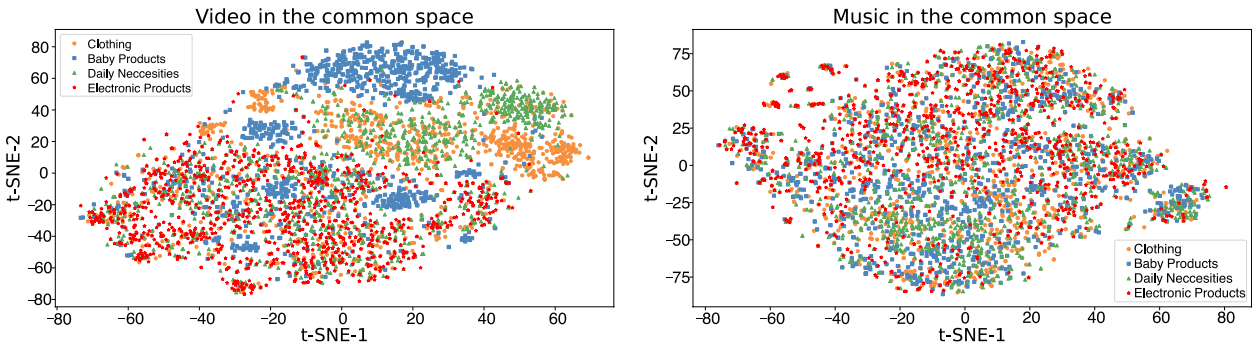
In the impact study of the WFM, we compare it with the traditional fusion methods, including Concat and Add.

We replace the WFM with Concat or Add respectively. For YuYin (Concat), the multi-modal features are concatenated and fed into the subsequent network. For YuYin (Add), due to the inconsistency of the feature dimensions, different features are first transformed to the same dimension by a linear projection and then summed up for fusion.

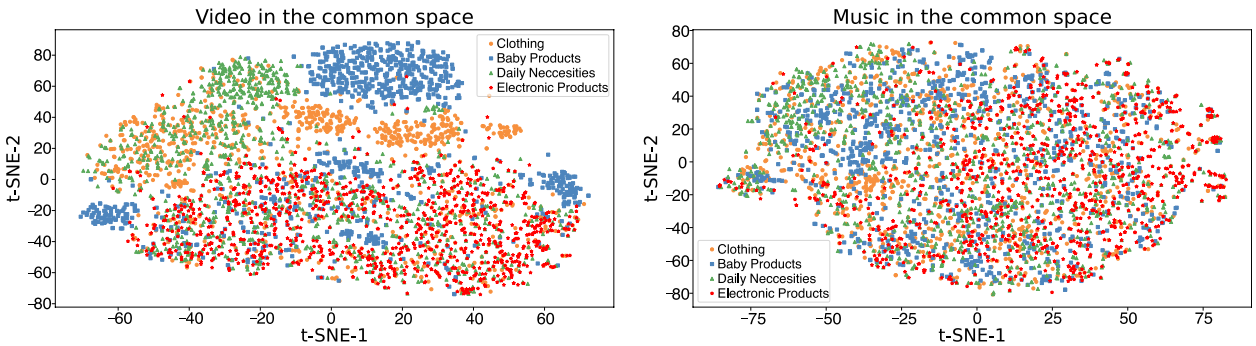
As shown in Table 4, YuYin (WFM) performs the best, which may be attributed to WFM refining the data processing granularity of the model by learning to weight different modalities in training, while the Concat and the Add method can hardly complete the targeted extraction of the data, which leads to more interference information in the fused data and limits the model performance. Furthermore, the results of YuYin (Add) may result from the missing information in the linear projection and summing up compared with direct concatenation in YuYin (Concat).

5.4.3 Effect of multi-task learning

To investigate the effect of multi-task learning, as shown in Fig. 6, kernel density estimation (KDE) is applied to visualize the similarity in video-music pairs to demonstrate how YuYin distinguishes between positive and negative video-music pairs in cross-matching. The results prove the similarity between positive video-music pairs is significantly bigger than that of negative pairs in the common space.



(a) The video and music projections from YuYin w/ prediction task.



(b) The video and music projections from YuYin w/o prediction task.

Fig. 7 Visualization of the video and music projections in the common space with dimension reduced to 2 by t-SNE

Furthermore, to explore the effect of the prediction task, as shown in Fig. 7, we observe the video and music features from YuYin with and without the prediction task, respectively. In detail, we use t-SNE to reduce the dimension and visualize the video and music features from each category in Commercial-98K. It shows that the video projections in the common space have a more distinct distribution with the prediction task. However, we also find there is no clear pattern in the distribution of the music features. The reason may owe to the lack of a fixed paradigm for selecting the background music, and even for the same advertisement, the music can be influenced by personal preferences, music popularity, and other factors.

6 Conclusion

To reduce the labor in manually selecting the background music for e-commerce advertisements. We first establish a large-scale dataset Commercial-98K from Alibaba, containing background music, videos, and product category tags of 98,071 advertisements. Then, we propose a video-music retrieval model YuYin with a novel WFM to fuse audio and emotion features and is trained by multi-task learning to cross-match video, music, and text as well as predict the category of video and music through a weight-shared classifier. We conduct experiments to find YuYin outperforms other models in video-music retrieval. and demonstrate the effect of multimodal and WFM in YuYin. Moreover, through visualization, we investigate the data distribution of each modality to prove YuYin can distinguish positive and negative video-music pairs in the common space. In the future, based on Commercial-98K, we will continue to carry out studies on the more effect factors besides emotion in video-music retrieval and replace our multi-modal feature extractors with the novel network.

Abbreviations

DNN	Deep neural network
CMR	Cross-modal retrieval
VMR	Video-music retrieval
WFM	Weighted fusion module
CNN	Convolutional neural network
PCA	Principal components analysis
NCE	Noise-contrastive rstimation
VAE	Variational auto-encoder
t-SNE	t-Distributed stochastic neighbor embedding

Acknowledgements
The authors acknowledge the support from the National Natural Science Foundation of China (No.62272409), the Key R &D Program of Zhejiang Province (No.2022C03126), and Project of Key Laboratory of Intelligent Processing Technology for Digital Music (Zhejiang Conservatory of Music), Ministry of Culture and Tourism (No.2022DMKLB001).

Authors' contributions
L. Ma and X. Wu conceived the study and implemented the method. R. Tang was responsible for data collection and processing, C. Zhong was involved in the technical design, L. Ma ran the experiment and wrote the first draft of the manuscript, and K. Zhang supervised the work and revised the first draft with L. Ma and X. Wu. All authors read and approved the final manuscript.

Availability of data and materials
The original datasets generated and/or analyzed during the current study are not publicly available due to copyright protection but are available from the corresponding author upon reasonable request.

Declarations

Competing interests
The authors declare that they have no competing interests.

Received: 15 June 2023 Accepted: 29 September 2023
Published online: 19 October 2023

References

1. J.I. Alpert, M.I. Alpert, Music influences on mood and purchase intentions. Psychol. Mark. **7**(2), 109–133 (1990)
2. G.C. Bruner, Music, mood, and marketing. J. Mark. **54**(4), 94–104 (1990)
3. J.I. Alpert, M.I. Alpert, Background music as an influence in consumer mood and advertising responses (ACR North American Advances, 1989)
4. N.B. Fernandez, W.J. Trost, P. Vuilleumier, Brain networks mediating the influence of background music on selective attention. Soc. Cogn. Affect. Neurosci. **14**(12), 1441–1452 (2019)
5. I. Salakka, A. Pitkäniemi, E. Pentikäinen, K. Mikkonen, P. Saari, P. Toiviainen, T. Särkämö, What makes music memorable? relationships between acoustic musical features and music-evoked emotions and memories in older adults. PLoS ONE **16**(5), e0251692 (2021)
6. F. Yi, J. Kang, Effect of background and foreground music on satisfaction, behavior, and emotional responses in public spaces of shopping malls. Appl. Acoust. **145**, 408–419 (2019)
7. K. Wang, Q. Yin, W. Wang, S. Wu, L. Wang, A comprehensive survey on cross-modal retrieval. arXiv preprint arXiv:1607.06215 (2016)
8. A. Zheng, M. Hu, B. Jiang, Y. Huang, Y. Yan, B. Luo, Adversarial-metric learning for audio-visual cross-modal matching. IEEE Trans. Multimed. **24**, 338–351 (2021)
9. Y. Liu, J. Wu, L. Qu, T. Gan, J. Yin and L. Nie, "Self-Supervised Correlation Learning for Cross-Modal Retrieval," in IEEE Transactions on Multimedia. **25**, 2851–2863 (2023) https://doi.org/10.1109/TMM.2022.3152086
10. A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, I. Sutskever, in *International Conference on Machine Learning* (PMLR, 2021), pp. 8821–8831
11. G. Wang, X. Xu, F. Shen, H. Lu, Y. Ji, H.T. Shen, Cross-modal dynamic networks for video moment retrieval with text query. IEEE Trans. Multimed. **24**, 1221–1232 (2022)
12. X. Song, J. Chen, Z. Wu and Y. -G. Jiang, "Spatial-Temporal Graphs for Cross-Modal Text2Video Retrieval," in IEEE Transactions on Multimedia. **24**, 2914–2923 (2022) https://doi.org/10.1109/TMM.2021.3090595
13. A. Radford, J.W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., in *International Conference on Machine Learning* (PMLR, 2021), pp. 8748–8763
14. Owens, A., & Efros, A. A. Audio-visual scene analysis with self-supervised multisensory features. In Proceedings of the European conference on computer vision (ECCV) (2018), pp. 631–648
15. Chen, L., Srivastava, S., Duan, Z., & Xu, C. Deep cross-modal audio-visual generation. In Proceedings of the on Thematic Workshops of ACM Multimedia 2017, (2017), pp. 349–357.
16. Nagrani, A., Albanie, S., & Zisserman, A. Seeing voices and hearing faces: Cross-modal biometric matching. In Proceedings of the IEEE conference on computer vision and pattern recognition. (2018) pp. 8427–8436.
17. R. Wang, H. Huang, X. Zhang, J. Ma, A. Zheng, in *2019 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)* (IEEE, 2019), pp. 300–305
18. Oh, T. H., Dekel, T., Kim, C., Mosseri, I., Freeman, W. T., Rubinstein, M., & Matusik, W. (2019). Speech2face: Learning the face behind a voice. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. (2019) pp. 7539–7548.
19. Chao, J., Wang, H., Zhou, W., Zhang, W., & Yu, Y. Tunesensor: A semantic-driven music recommendation service for digital photo albums. In

Proceedings of the 10th International Semantic Web Conference. (2011). ISWC2011 (October 2011).
20. Y. Jia, L. Bai, S. Liu, P. Wang, J. Guo, Y. Xie, Semantically-enhanced kernel canonical correlation analysis: a multi-label cross-modal retrieval. Multimedia Tools Appl. **78**(10), 13169–13188 (2019)
21. Hong, S., Im, W., & Yang, H. S. Cbvmr: content-based video-music retrieval using soft intramodal structure constraint. In Proceedings of the 2018 ACM on international conference on multimedia retrieval. (2018) pp. 353–361
22. Li, B., & Kumar, A. Query by Video: Cross-modal Music Retrieval. In ISMIR (2019) pp. 604–611
23. Surís, D., Duarte, A., Salvador, A., Torres, J., & Giró-i-Nieto, X. Cross-modal embeddings for video and audio retrieval. In Proceedings of the european conference on computer vision (eccv) workshops (2018) pp. 0-0
24. B. Dai, The impact of online shopping experience on risk perceptions and online purchase intentions: the moderating role of product category and gender. Ph.D. thesis (2007)
25. R. Jain, S. Bagdare, Music and consumption experience: a review. Int. J. Retail Distrib. Manag. **39**(4), 289–302 (2011)
26. M.F. Zander, Musical influences in advertising: How music modifies first impressions of product endorsers and brands. Psychol. Music **34**(4), 465–480 (2006)
27. S. Koelstra, Deap: A database for emotion analysis; using physiological signals. IEEE Trans. Affect. Comput. **3**(1), 18–31 (2012)
28. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C. L., & Parikh, D. Vqa: Visual question answering. In Proceedings of the IEEE international conference on computer vision (2015) pp. 2425–2433
29. Li, T., Sun, Z., Zhang, H., Li, J., Wu, Z., Zhan, H., ... & Shi, H. Deep music retrieval for finegrained videos by exploiting cross-modal-encoded voice-overs. In Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (2021) pp. 1880–1884
30. C.L. Liu, Y.C. Chen, Background music recommendation based on latent factors and moods. Knowl.-Based Syst. **159**, 158–170 (2018)
31. Zhou, L., Xu, C., & Corso, J. Towards automatic learning of procedures from web instructional videos. In Proceedings of the AAAI Conference on Artificial Intelligence. **32**(1), (2018)
32. H.T.P. Thao, G. Roig, D. Herremans, EmoMV: Affective music-video correspondence learning datasets for classification and retrieval. Inf. Fusion **91**, 64–79 (2023). https://doi.org/10.1016/j.inffus.2022.10.002
33. Xu, J., Mei, T., Yao, T., & Rui, Y. Msr-vtt: A large video description dataset for bridging video and language. In Proceedings of the IEEE conference on computer vision and pattern recognition (2016) pp. 5288–5296
34. J. Yi, Y. Zhu, J. Xie and Z. Chen, "Cross-Modal Variational Auto-Encoder for Content-Based Micro-Video Background Music Recommendation," in IEEE Transactions on Multimedia. **25**, 515–528 (2023). https://doi.org/10.1109/TMM.2021.3128254.
35. S. Abu-El-Haija, N. Kothari, J. Lee, P. Natsev, G. Toderici, B. Varadarajan, S. Vijayanarasimhan, Youtube-8m: A large-scale video classification benchmark. arXiv preprint arXiv:1609.08675 (2016)
36. Plummer, B. A., Wang, L., Cervantes, C. M., Caicedo, J. C., Hockenmaier, J., & Lazebnik, S. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to sentence models. In Proceedings of the IEEE international conference on computer vision (2015) pp. 2641–2649
37. Miech, A., Zhukov, D., Alayrac, J. B., Tapaswi, M., Laptev, I., & Sivic, J. Howto100m: Learning a textvideo embedding by watching hundred million narrated video clips. In Proceedings of the IEEE/CVF international conference on computer vision (2019) pp. 2630–2640.
38. D.R. Hardoon, S. Szedmak, J. Shawe-Taylor, Canonical correlation analysis: An overview with application to learning methods. Neural Comput. **16**(12), 2639–2664 (2004)
39. G.F. Zebende, Dcca cross-correlation coefficient: Quantifying level of cross-correlation. Phys. A Stat. Mech. Appl. **390**(4), 614–618 (2011)
40. Wu, X., Qiao, Y., Wang, X., & Tang, X. Cross matching of music and image. In Proceedings of the 20th ACM international conference on Multimedia (2012) pp. 837–840
41. Z. Wang, Y. Li, R. Hong, X. Tian, Eigenvector-based distance metric learning for image classification and retrieval. ACM Trans. Multimed. Comput. Commun. Appl. (TOMM) **15**(3), 1–19 (2019)
42. L. Shen, R. Hong, H. Zhang, X. Tian, M. Wang, Video retrieval with similarity-preserving deep temporal hashing. ACM Trans. Multimed. Comput. Commun. Appl. (TOMM) **15**(4), 1–16 (2019)
43. L. Zhang, H. Guo, K. Zhu, H. Qiao, G. Huang, S. Zhang, H. Zhang, J. Sun, J. Wang, Hybrid modality metric learning for visible-infrared person re-identification. ACM Trans. Multimed. Comput. Commun. Appl. (TOMM) **18**(1s), 1–15 (2022)
44. R. Cao, Q. Zhang, J. Zhu, Q. Li, Q. Li, B. Liu, G. Qiu, Enhancing remote sensing image retrieval using a triplet deep metric learning network. Int. J. Remote Sens. **41**(2), 740–751 (2020)
45. J. Wei, Y. Yang, X. Xu, X. Zhu, H.T. Shen, Universal weighting metric learning for cross-modal retrieval. IEEE Trans. Pattern. Anal. Mach. Intell. **44**(10), 6534–6545 (2021)
46. X. Gu, Y. Shen, C. Lv, A dual-path cross-modal network for video-music retrieval. Sensors **23**(2), 805 (2023)
47. L. Pretet, G. Richard, C. Souchier, G. Peeters, Video-to-Music Recommendation using Temporal Alignment of Segments. IEEE Trans. Multimed. 1 (2022). https://doi.org/10.1109/TMM.2022.3152598
48. Szegedy, C., Ioffe, S., Vanhoucke, V., & Alemi, A. Inception-v4, inception-resnet and the impact of residual connections on learning. In Proceedings of the AAAI conference on artificial intelligence **31**(1), (2017)
49. S. Hershey, S. Chaudhuri, D.P.W. Ellis, J.F. Gemmeke, A. Jansen, C. Moore, M. Plakal, D. Platt, R.A. Saurous, B. Seybold, M. Slaney, R. Weiss, K. Wilson, in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2017). https://arxiv.org/abs/1609.09430
50. Eyben, F., Wöllmer, M., & Schuller, B. Opensmile: the munich versatile and fast open-source audio feature extractor. In Proceedings of the 18th ACM international conference on Multimedia (2010) pp. 1459–1462
51. Mignon, A., & Jurie, F. CMML: A new metric learning approach for cross modal matching. In Asian Conference on Computer Vision (2012) pp. 14-pages
52. Y. Wu, S. Wang, G. Song, Q. Huang, Online asymmetric metric learning with multi-layer similarity aggregation for cross-modal retrieval. IEEE Trans. Image Process. **28**(9), 4299–4312 (2019)
53. M. Gutmann, A. Hyvärinen, in *Proceedings of the thirteenth international conference on artificial intelligence and statistics* (JMLR Workshop and Conference Proceedings, 2010), pp. 297–304
54. Ge, W. Deep metric learning with hierarchical triplet loss. In Proceedings of the European conference on computer vision (ECCV) (2018) pp. 269–285
55. Y. Zhou, Z. Wang, C. Fang, T. Bui, T.L. Berg, in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2018), pp. 3550–3558. https://doi.org/10.1109/CVPR.2018.00374
56. Pons, J., & Serra, X. musicnn: Pre-trained convolutional neural networks for music audio tagging. arXiv preprint arXiv:1909.06654 (2019)
57. Y.Y. Yang, M. Hira, Z. Ni, A. Astafurov, C. Chen, C. Puhrsch, D. Pollack, D. Genzel, D. Greenberg, E.Z. Yang, et al., in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, 2022), pp. 6982–6986
58. Y. Gong, Y.A. Chung, J. Glass, Ast: Audio spectrogram transformer. arXiv preprint arXiv:2104.01778 (2021)
59. Y. Cui, W. Che, T. Liu, B. Qin, Z. Yang, Pre-training with whole word masking for Chinese Bert. IEEE/ACM Trans. Audio Speech Lang. Process. **29**, 3504–3514 (2021)
60. J.B. Alayrac, A. Recasens, R. Schneider, R. Arandjelović, J. Ramapuram, J. De Fauw, L. Smaira, S. Dieleman, A. Zisserman, Self-supervised multimodal versatile networks. Adv. Neural Inf. Process. Syst. **33**, 25–37 (2020)
61. G. Andrew, R. Arora, J. Bilmes, K. Livescu, in *International conference on machine learning* (PMLR, 2013), pp. 1247–1255

Publisher's Note