Continuous Emotion-Based Image-to-Music Generation

Yajie Wang, Mulin Chen, and Xuelong Li, *Fellow, IEEE*

***Abstract*—Image-to-music generation aims to generate realistic pure music according to a given image. Although many previous works are conducted on bridging image and music, they mainly focus on the content-based cross-modal matching. For example, matching the Christmas song to an image that contains a Christmas tree. By comparison, image-to-music generation is a more challenging task due to its ambiguity and subjectivity. Specifically, there is no explicit correlation between the image content and music melody, without any lyric and human sound. Meanwhile, the perception of generated music varies from person to person. Inspired by the synesthesia phenomenon, we think that if an image tends to elicit a certain emotion on human, the generated music should also leave a similar impression. Therefore, in this paper, we propose a continuous emotion-based image-to-music generation framework, which uses emotion as the key for cross-modal generation. Specifically, a new image-music dataset is established, which uses valence-arousal (VA) space to capture the complex and nuanced nature of emotions. After that, a plug and play model is proposed to translate an image into a piece of music with similar emotion, which projects the emotions into continuous-valued labels, and explores both the intra-modal and inter-modal emotional consistency with contrastive learning. To our best knowledge, this is the first end-to-end framework towards the task of pure music generation from natural images. Extensive experiments show that the generated music achieves satisfactory emotional consistency with the input images, as well as impressive quality.**

***Index Terms*—Image-to-music generation, valence-arousal space, multi-modal cognitive computing, vicinagearth security.**

I. Introduction

In daily life, when people see an image, they often associate it with a piece of melody. This phenomenon has been extensively studied in the field of cognition and is referred to as synesthesia. Researches on synesthesia have shown that stimulation of one sensory channel may trigger a response in another channel, leading to a complex and integrated sensation. The above theory has been extensively applied in real-world applications. In artistic creation, Franz Liszt's "Les Preludes" uses contrasting dynamics and tempos to create an emotional visual imagery of a stormy night for the listeners. In music therapy, Guided Imagery and Music (GIM) [1] is a therapeutic approach that employs music to induce a state of relaxed awareness, where individuals can revisit previous pleasant visual imagery and achieve therapeutic benefits. Therefore, achieving cross-modal translation between images and music is a task of high significance. However, it remains to be a challenging problem in the field of artificial intelligence. In this paper, we focus on generating pure music from a given image.

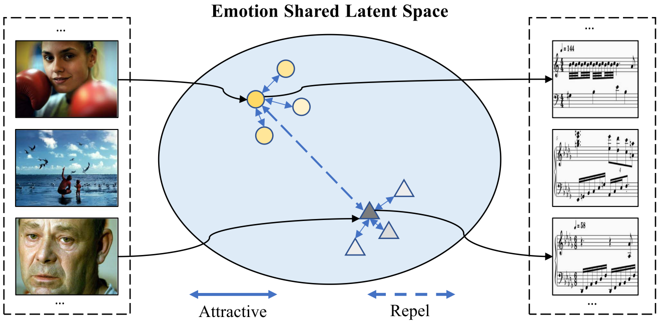

Fig. 1. Basic idea of our emotionally consistent music generation for a given image in continuous valence-arousal space.

Generating pure music from images is a challenging task due to its **ambiguity** and **subjectivity**. 1) The **ambiguity** arises from the fact that the correlation between the image content and music melody is not explicit, compared with image-to-text generation. Thus, some studies in image-music matching [2] and image-to-song generation [3] use lyrics as an intermediary to connect the image and music modalities, rather than directly seeking correspondence between the two modalities themselves. 2) The **subjectivity** in this process comes from the fact that different individuals have varying interpretations of the details in music, which results in a lack of uniform evaluation standards for music generation and makes it difficult to optimize the model. According to the synesthesia theory, what links music and images together is emotion, rather than low-level features [4]. Therefore, we use emotion as an intermediate bridge between image and music. Although emotion itself may be subject to individual perception, it also demonstrates universal patterns. For example, when people listen to Mozart's Turkish March, they tend to experience a feeling of happiness, which can be considered as a suitable standard for most individuals. Therefore,

Manuscript received 20 July 2023; revised 26 October 2023; accepted 16 November 2023. Date of publication 6 December 2023; date of current version 4 April 2024. This work was supported in part by the National Key Research and Development Program of China under Grant 2022ZD0160803, and in part by the National Natural Science Foundation of China under Grants 61871470 and 62106182. The Associate Editor coordinating the review of this manuscript and approving it for publication was Mr. J. Song. *(Mulin Chen and Xuelong Li contributed equally to this work.) (Corresponding authors: Xuelong Li; Mulin Chen.)*

The authors are with the School of Artificial Intelligence, OPtics and ElectroNics (iOPEN), Northwestern Polytechnical University, Xi'an 710072, China, and also with the Key Laboratory of Intelligent Interaction and Applications, Ministry of Industry and Information Technology, Northwestern Polytechnical University, Xi'an 710072, China (e-mail: yajie_wang@mail.nwpu.edu.cn; chenmulin@nwpu.edu.cn; li@nwpu.edu.cn).

Digital Object Identifier 10.1109/TMM.2023.3338089

1520-9210 © 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: Hanyang University. Downloaded on December 20,2024 at 07:33:50 UTC from IEEE Xplore. Restrictions apply.

emotion can still serve as a criterion that caters to most people. In our task, assuming that if an image can evoke a certain emotion in a person, the generated music should evoke a similar sensation, this can serve as a metric for evaluating the quality of pure music generation. By using emotion as a medium, issues of ambiguity and subjectivity can be alleviated.

Meanwhile, as a new task, the generation of music from images imposes new requirements on the data used for training. 1) The music dataset is better to be controllable, as generating music with structure requires a high level control over music elements such as note. In all the existing cross-modal datasets, music is in audio format. This spectral representation of music is with less controllability as it is an analogue signal that cannot be separated into notes. 2) The emotional labels for cross-modal datasets are better suited to be continuous. Discrete emotional approaches are effective in a variety of tasks, such as emotion recognition and affective computing. However, in the proposed task, the searching space of music notes is very large, and fine-grained emotion is required to distinguish the subtle emotional changes within images, thus generate more accurate music accordingly. Therefore, continuous emotion space is more appropriate for our task. To sum up, it is urgent to establish an appropriate dataset for emotion-based image-to-music generation.

To address the aforementioned issues, we construct a new image-music dataset with continuous emotion labels, and propose a continuous emotion-based image-to-music generation framework. Specifically, continuous valence-arousal (VA) emotion model is used as the key for cross-modal generation to tackle ambiguity and subjective issues, while better captures the nuance of emotions in images compared to categories. VA emotion is a two-dimension space that represents the valence-arousal dimensions of emotions, where valence refers to the positivity or negativity of an emotion, ranging from negative to positive while arousal refers to the intensity or activation level of an emotion, ranging from low to high. Based on the VA space, a cross-modal image-music dataset is established, which contains MIDI-format music dataset. MIDI-format music is in the form of digital data, and offers more precise control and editing flexibility compared to audio. After that, a plug and play model is proposed to translate an image into a piece of music with similar emotion, which projects the emotions into continuous-valued labels, and explores both the intra-modal and inter-modal emotional consistency with contrastive learning. The basic idea of the proposed method is shown in Fig. 1. Our code and dataset are available at https://github.com/zBaymax/EIMG.

In summary, the contributions of this paper are threefold:

1) The first end-to-end framework is proposed for generating pure music from natural images directly, without resorting to image captions or lyrics. Considering the ambiguity and subjectivity of the task itself, emotion is introduced as a medium to guide the cross-modal translation procedure.
2) A plug and play model is put forward to translate an image into a piece of music with contrastive learning. It reduces the distance between images and music with similar emotions, as well as between images or music within the same modality, which is effective to process continuous-valued labels.
3) A new image-music dataset is constructed for emotion-based cross-modal generation. It contains MIDI-format music, resulting in precise control and manipulation of music elements. Moreover, the use of fine-grained continuous emotion labels allows for a more comprehensive representation of the complex and subtle human emotions.

II. Related Work

In this section, a brief introduction to emotion representation models is provided, and then we will discuss relevant works on music generation. Finally, related works on image-to-music generation are introduced.

A. *Emotion Representation Models*

There are two typical models for representing emotions in psychological: Categorical Emotion States (CES) and Dimensional Emotion Space (DES). CES models posit that emotions are discrete and can be classified into a limited number of basic categories, such as Mikel's eight emotions [11] and Ekman's six emotions [12]. However, CES models have been criticized for oversimplifying the complexity of emotions and for not capturing the individual differences and nuances of emotional experiences. DES model, on the other hand, assume that emotions are more complex and multidimensional, and each emotion can be represented as a point in a continuous space of emotional dimensions. The valence-arousal-dominance (VAD) emotion model is representative of the DES model, where valence refers to the positivity or negativity of an emotion, ranging from negative to positive while arousal refers to the intensity or activation level of an emotion, ranging from low to high and dominance refers to the level of control or power associated with an emotion, ranging from submissive to dominant. This model has been developed to address the limitations of the CES model and to provide a more nuanced and comprehensive understanding of emotions. As valence and arousal are considered the most critical dimensions in emotional experiences, the difficulty in predicting dominance has led many studies [13], [14], [15] to represent emotions in the valence-arousal (VA) space instead. Some works [16], [17], [18], [19], [20] have employed emotion for aligning different modalities. However, these works are mostly limited to retrieval [16], matching [17], [18], or simple concatenation of existing music to construct longer sequences [19], [20]. In this paper, continuous VA emotion is introduced to express the complexity of emotions for the cross-modal translation procedure.

B. *Music Generation*

For symbolic music generation, some models [23], [24], [25] use generative adversarial network to maximum likelihood training and generate new pieces. However, generating sequences with GANs can be challenging due to the instability of training and mode collapse. Since music can be represented as sequences, autoencoders [26], [27], [28] and transformers [29], [30], [31], [32] are gradually applied in music generation. In addition to the generation of symbolic music, some models [33], [34] generate music by analyzing audio waveforms. Our model

Authorized licensed use limited to: Hanyang University. Downloaded on December 20,2024 at 07:33:50 UTC from IEEE Xplore. Restrictions apply.

employs autoencoders for music generation based on the image emotions.

C. Image-to-Music Generation

BGT [3] converts the task into a three-stage process of image-to-caption, caption-to-lyrics, and lyrics-to-music generation, with three existing models used for each stage of the task. Others involve directly mapping low-level features of the image, such as colors [35], [36] and edges [35], to corresponding musical notes and then generating music by concatenating these notes. However, the music generated by directly mapping low-level image features to musical notes may lack overall structure or musicality, and these models do not directly justify the perceived similarity between a pair of image and pure music, resulting in a lack of accuracy and fidelity in the final music. In the related field of video background music generation, recent studies [37], [38] attempts to explore the connection between video and music by matching their three respective features, and achieves high performance. However, they rely solely on low-level features and limit to access more complex and meaningful semantic features. Meanwhile, these studies [37], [38], [39] primarily concentrate on rhythm transformations in videos, such as motion, but lack the emotional analysis specific to the content itself, which differs from the focus of image-to-music generation task. In this paper, emotion is introduced to justify the similarity between a pair of image and pure music for the cross-modal translation procedure. Further, a plug and play model is proposed for the task of emotion-based image-to-music generation, with contrastive learning exploring emotional consistency both within and across modalities.

III. Establish Cross-Modal Image-Music Dataset

In this section, a cross-modal image-music dataset is constructed for emotion-based cross-modal generation. The MIDI piano music dataset with VA values is proposed to enable precise control and manipulation of music elements. The VA model maps emotions on 2D space based on valence and arousal, where valence is positivity/negativity and arousal is intensity. Then, the emotion matching score is introduced to measure the emotional correlation between proposed music dataset and two existing image datasets. Finally, a cross-modal image-music dataset is constructed to capture the complex and nuanced nature of emotions.

A. Music Data Preparation

The music dataset is a collection of 467 piano pieces that are manually segmented into 3000 music clips of 15 seconds each for emotional annotation. The collection are gathered from the internet, and features a diverse range of musical genres, including movie soundtracks, classical music, pop music, and Japanese anime music. In the aspect of music labels, continuous valence-arousal emotion representation model is considered to express people's complex emotions.

For dataset annotation, six annotators were selected, consisting of three males and three females, with two having a background in music theory. This approach aims to consider the potential impact of music theory on emotional perception while also controlling for gender effects. After completing the audition, the annotators were assigned VA values ranging from 1 to 9 based on the emotions they perceived in the music. All six annotators worked independently without interference, while maintaining close communication with the organizers to ensure a shared standard for valence-arousal values. The final VA annotation value is the average of the annotations from the six annotators. After completing the annotation, an objective analysis of the results is conducted, as shown in Table I. The data in the table shows that higher arousal values, which indicate higher intensity of emotional expression, correspond to denser musical notes (1.710, 1.720), and lower arousal values correspond to sparser musical notes (1.044, 1.286). This trend applies to both happy (high-valence) and sad (low-valence) emotions. Additionally, higher valence values also correspond to denser musical notes at the same level of arousal. Therefore, our annotations are relatively valid and reliable. The comparison between proposed and existing music datasets with VA values is summarized in Table II.

TABLE I
Analysis Between the Density of Musical Notes and the Annotation Results in the Music Dataset

	Low-Valence	High-Valence
High-Arousal	1.710	1.720
Low-Arousal	1.044	1.286

TABLE II
Comparison Between Our Released Music Dataset and Other Existing Public Datasets With VA Values

Name	Label Type	Size	Modality
EMOPIA [5]	Russell's 4Q	1078	MIDI
DEAM [6]	VA values	1802	Audio
Emo-Soundscapes [7]	VA values	1213	Audio
CCMED-WCMED [8]	VA values	800	Audio
Emusic [9]	VA values	140	Audio
VGMIDI [10]	VA values	95	MIDI
Ours	VA values	3000	MIDI

B. Image-Music Data Preparation

Our image dataset is constructed by combining the International Affective Picture System (IAPS) [40] and the Nencki Affective Picture System (NAPS) [41], which are both labelled with VA values. The IAPS is a database of pictures designed to provide a standardized set of pictures for studying emotion, which is widely used in psychological research. The NAPS consists of 1,356 realistic, high-quality photographs that are divided into five categories (people, faces, animals, objects, and landscapes), in which pictures were rated according to the valence, arousal, and approach-avoidance dimensions using computerized bipolar semantic slider scales.

Due to the different scales of the image and music labels,the VA values of each dataset are normalized into the range of [1,9], respectively. Then, the image and music datasets with

Authorized licensed use limited to: Hanyang University. Downloaded on December 20,2024 at 07:33:50 UTC from IEEE Xplore. Restrictions apply.

TABLE III
COMPARISON OF OUR RELEASED IMAGE-MUSIC DATASET WITH OTHER EXISTING PUBLIC DATASETS

Name	Image Size	Music Size	Construction Method	Music Length	Music Modality
IMAC [21]	85000	3812	Emotion Category	60s	Audio
Music-Image [22]	–	22632	Music Video Cut	60s	Audio
Shuttersong [2]	–	586	Lyric Alignment	Undefined	Audio
IMEMNET [14]	26620	1820	VA Distance	2s	Audio
Ours	2538	3000	VA Distance	15s	MIDI

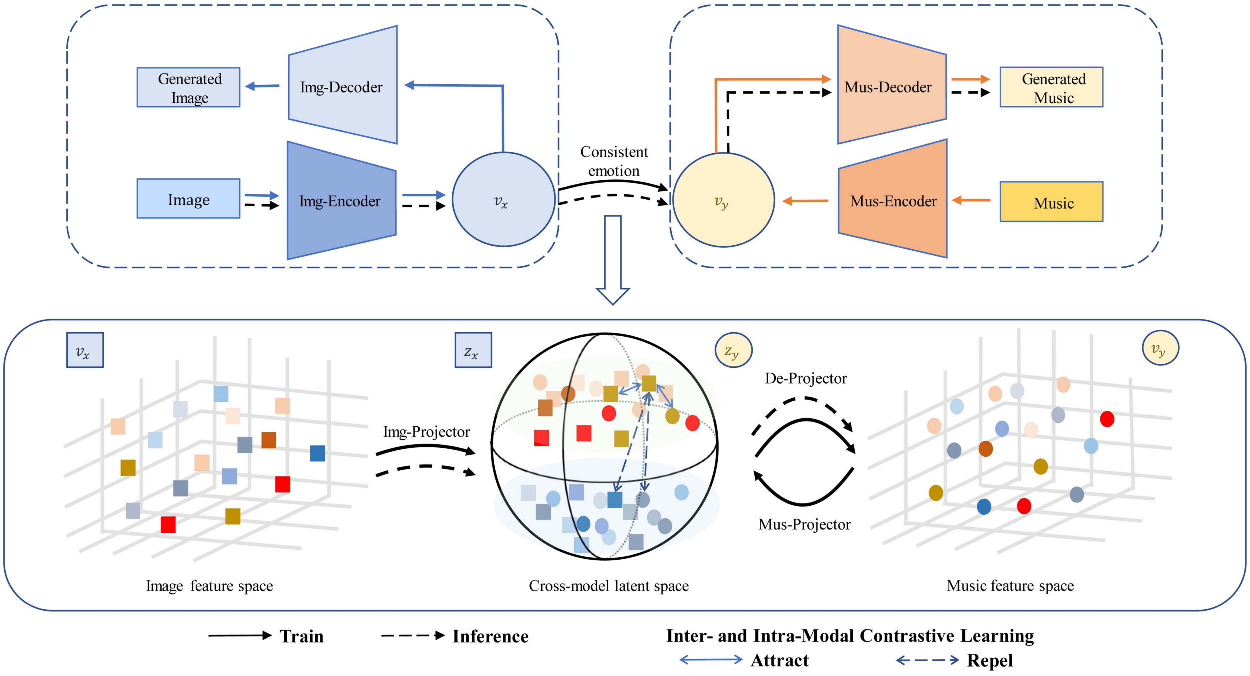

Fig. 2. Overview of the proposed continuous emotion-based image-to-music generation. An end-to-end model is proposed to translate an image into a piece of music with similar emotion, which projects the emotions into continuous-valued labels, and explores both the intra-modal and inter-modal emotional consistency with contrastive learning. During inference, the feature vector extracted from the image by the image encoder is transformed into a music feature vector in the cross-modal latent space, and then passed through the music decoder to generate music that is emotionally consistent with the image.

continuous VA values are combined to construct cross-modal image-music dataset by setting the formula for calculating the emotion matching score as

$$\mathrm{sim}(\mathbf{x},\mathbf{y}) = \left[(v_x - v_y)^2 + (a_x - a_y)^2\right]^{-\frac{1}{2}}, \tag{1}$$

where $\mathrm{sim}(\mathbf{x},\mathbf{y})$ represents the emotion similarity score between image $\mathbf{x}$ and music $\mathbf{y}$. v and a are two continuous dimensions used to represent emotions, which represent the valence and arousal values respectively. The valence dimension ranges from unpleasant to pleasant, and the arousal dimension ranges from calm to excited.

Equation (1) calculates the emotion similarity score between the valence-arousal coordinates of the image and the music. A larger value indicates a closer emotion matching between the two. By calculating emotion similarity scores between each image and each music in the dataset, a comprehensive cross-modal image-music dataset is constructed that captures the fine-grained emotional similarity between the two modalities. The comparison of proposed dataset with existing cross-modal image-music datasets is summarized in Table III.

IV. EMOTION-BASED IMAGE-TO-MUSIC GENERATION

In this section, an end-to-end model is introduced to capture the emotion of an image, and generate a piece of music according to the learned emotion. The proposed model employs two autoencoders to separately map images and music into two distinct feature spaces, which preserve the modality-specific features. After that, a shared cross-modal latent space is established to capture the emotional state. Since emotion-based images and music have a many-to-many correspondence, and share consistent features for the same emotion within each modality, a new contrastive learning strategy for continuous labels is designed to learn the consistency of emotions within and across modalities in the cross-modal latent space.

A. Network Structure

The overview of the model is shown as Fig. 2. Specifically, we initially train the image autoencoder (Img-Encoder and Img-Decoder) and the music autoencoder (Mus-Encoder and Mus-Decoder) to construct feature spaces within each modality. Subsequently, the image $\mathbf{x}$ is encoded by Img-Encoder to

Authorized licensed use limited to: Hanyang University. Downloaded on December 20,2024 at 07:33:50 UTC from IEEE Xplore. Restrictions apply.

obtain $\mathbf{v_x}$, and the music $\mathbf{y}$ is encoded by Mus-Encoder to obtain $\mathbf{v_y}$. These vectors preserve the modality-specific features. Then, with two projectors (Img-Projector and Mus-Projector), they are projected into the emotion shared cross-modal latent space separately, yielding the embeddings $\mathbf{z_x}$ and $\mathbf{z_y}$. The formulas of this process are defined as

$$\begin{aligned}\mathbf{v_x} &= \text{Img-Encoder}(\mathbf{x}),\\ \mathbf{v_y} &= \text{Mus-Encoder}(\mathbf{y}),\\ \mathbf{z_x} &= \text{Img-Projector}(\mathbf{v_x}),\\ \mathbf{z_y} &= \text{Mus-Projector}(\mathbf{v_y}),\\ \tilde{\mathbf{v}}_\mathbf{y} &= \text{De-Projector}(\mathbf{z_y}).\end{aligned} \tag{2}$$

Then, reconstruction loss is designed to ensure that the features projected into the latent space can be reconstructed back to the original feature space for music generation. The calculation formula for the reconstruction loss is as

$$L_{rec} = \|\tilde{\mathbf{v}}_\mathbf{y} - \mathbf{v_y}\|_2. \tag{3}$$

Furthermore, in the emotion shared cross-modal latent space, contrastive learning (introduced in Section IV-B) is proposed to explore the emotion consistency between two modalities and within single modality itself, making the music and images with similar emotions closer and those with different emotions farther away.

During the inference phase, the image feature is projected into the emotion shared cross-modal latent space to obtain $\mathbf{z_x}$, which is further projected back to the music feature space and utilize Mus-Decoder to generate music. The proposed model is compatible with most image autoencoders and music autoencoders and has plug-and-play characteristic.

B. Cross-Modal Continuous Label Supervised Contrastive Learning

In order to explore emotional consistency within latent space, a contrastive learning is proposed in this part. The valence-arousal (VA) model represents the complex emotions of humans in a 2D Cartesian space, where valence represents the degree of pleasantness, and arousal indicates the intensity of the emotion. The closer the VA values of an image and a piece of music are, the more similar the emotions evoked by the two stimuli. However, the emotional correspondence is not one-to-one relationships, as there are many music pieces that correspond to the same emotion as a given image. For example, a happy image can be associated with many different types of happy music varied greatly in note pitch and melody. Nonetheless, studies [5], [42] also show that music with similar emotion tends to exhibit certain commonalities in pitch, melody, and other acoustic features, despite not being identical. For instance, emotionally intense music generally has a higher note density, while soothing music typically corresponds to a lower note density. Similar rules also apply in the image modality. Therefore, in order to generate diverse music that is emotionally consistent with images, the proposed contrastive learning strategy aims to learn both the intra-modal emotion consistency and the inter-modal emotion correlation, which preserves the emotion-related information within each modality, and transforms it across different modalities effectively. The proposed contrastive learning on continuous labels is illustrated in Fig. 3.

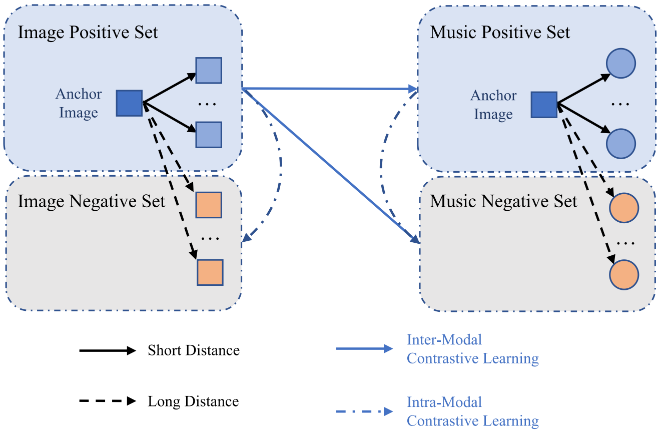

Fig. 3. Proposed inter-modal and intra-modal contrastive learning on continuous labels.

Since there can be multiple pieces of music that correspond to the same emotion of a given music, we group the music pieces and images based on their VA values. This allows us to learn more generalized emotional correlations between different music and images. Specifically, an image is selected at randomly from the dataset to serve as the anchor image. Given the image anchor $\mathbf{x}^*$, to construct the music positive set $\{\mathbf{y}\}_+$ and music negative set $\{\mathbf{y}\}_-$, we select the top n music clips with the highest emotion similarity scores to $\mathbf{x}^*$, and the bottom n music clips with the lowest emotion similarity scores to $\mathbf{x}^*$, respectively, from the music dataset. The formulas of this process are as

$$\begin{aligned}\{\mathbf{y}\}_+ &= \{\mathbf{y_i}|\text{sim}(\mathbf{y_i}, \mathbf{x}^*) \in T_n(\text{sim}(\mathbf{y_i}, \mathbf{x}^*), \mathbf{y_i} \in \mathbf{M})\},\\ \{\mathbf{y}\}_- &= \{\mathbf{y_i}|\text{sim}(\mathbf{y_i}, \mathbf{x}^*) \in B_n(\text{sim}(\mathbf{y_i}, \mathbf{x}^*), \mathbf{y_i} \in \mathbf{M})\},\end{aligned} \tag{4}$$

where $\mathbf{M}$ is the set of all music clips, $\mathbf{x}^*$ is the image anchor, and T_n and B_n represent the top n and bottom n elements with the highest and lowest similarity scores, respectively. Given the potential for multiple images corresponding to the same emotion, we commence by randomly selecting an image $\mathbf{x}^*$ from the dataset as the anchor. Subsequently, we construct the image positive set (denoted as $\{\mathbf{x}\}_+$) and the image negative set (denoted as $\{\mathbf{x}\}_-$) by selecting n images with the highest and lowest emotion similarity scores to the anchor image $\mathbf{x}^*$ from the image dataset. The formulas of this process are as

$$\begin{aligned}\{\mathbf{x}\}_+ &= \{\mathbf{x_i}|\text{sim}(\mathbf{x_i}, \mathbf{x}^*) \in T_n(\text{sim}(\mathbf{x_i}, \mathbf{x}^*), \mathbf{x_i} \in \mathbf{I})\},\\ \{\mathbf{x}\}_- &= \{\mathbf{x_i}|\text{sim}(\mathbf{x_i}, \mathbf{x}^*) \in B_n(\text{sim}(\mathbf{x_i}, \mathbf{x}^*), \mathbf{x_i} \in \mathbf{I})\},\end{aligned} \tag{5}$$

where $\mathbf{I}$ is the set of all images, $\mathbf{x}^*$ is the image anchor, and T_n and B_n represent the top n and bottom n elements with the highest and lowest emotion similarity scores, respectively.

Once we have obtained the positive ($\{\mathbf{x}\}_+$, $\{\mathbf{y}\}_+$) and negative sets ($\{\mathbf{x}\}_-$, $\{\mathbf{y}\}_-$) of image and music respectively, we can use them to perform inter-modal and intra-modal contrastive

Authorized licensed use limited to: Hanyang University. Downloaded on December 20,2024 at 07:33:50 UTC from IEEE Xplore. Restrictions apply.

learning. Inter-modal contrastive learning explores cross-modal emotion consistency and intra-modal contrastive learning allows us to learn more robust and discriminative latent vectors in cross-modal latent space for each modality.

Inter-Modal Contrastive Learning: Given the image positive set $\{\mathbf{x}\}_+$ as the anchor, the music positive set $\{\mathbf{y}\}_+$ as the positive sample, and the music negative set $\{\mathbf{y}\}_-$ as the negative sample, we define the score function between each image $\mathbf{x}$ and music $\mathbf{y}$ as

$$\mathrm{S}_{\mathrm{inter}}(\mathbf{x}, \mathbf{y}) = \cos(\mathbf{z}_{\mathbf{x}}, \mathbf{z}_{\mathbf{y}})/\tau, \tag{6}$$

where $\cos(\mathbf{u}, \mathbf{v}) = \mathbf{u}^{\mathrm{T}}\mathbf{v}/\|\mathbf{u}\|\mathbf{v}\|$ denotes cosine similarity, and τ denotes a temperature hyper-parameter. $\mathbf{z}_{\mathbf{x}}$ is obtained by projecting the feature vector $\mathbf{v}_{\mathbf{x}}$, extracted from image $\mathbf{x}$ using the Img-Encoder, into the cross-modal latent space. Similarly, $\mathbf{z}_{\mathbf{y}}$ is obtained by projecting the feature vector $\mathbf{v}_{\mathbf{y}}$, extracted from music $\mathbf{y}$ using the Mus-Encoder, into the cross-modal latent space. The contrastive loss between image positive set $\{\mathbf{x}\}_+$, music positive set $\{\mathbf{y}\}_+$ and music negative set $\{\mathbf{y}\}_-$ is computed as

$$L_{inter} = \sum_{\mathbf{x} \in \{\mathbf{x}\}_+} -\frac{1}{n} \sum_{\mathbf{y} \in \{\mathbf{y}\}_+} \log \frac{\exp(\mathrm{S}_{\mathrm{inter}}(\mathbf{x}, \mathbf{y}))}{\sum_{\mathbf{a} \in \{\{\mathbf{y}\}_+, \{\mathbf{y}\}_-\}} \exp(\mathrm{S}_{\mathrm{inter}}(\mathbf{x}, \mathbf{a}))}, \tag{7}$$

where n represents the number of elements in set $\{\mathbf{y}\}_+$. This form of contrastive loss is inspired by supervised contrastive loss [43]. By using the proposed inter-modal contrastive loss, we can close the distance between images and music that share similar emotion, while extending the distance between those that have dissimilar emotion. This allows us to better capture emotion relationships between images and music, and generate cross-modal content that is emotionally consistent.

Intra-Modal Contrastive Learning: For image modality, given image positive set $\{\mathbf{x}\}_+$ and image negative set $\{\mathbf{x}\}_-$, we define the score function between two images as

$$\mathrm{S}_{\mathrm{inter}_{\mathrm{x}}}(\mathbf{x_1}, \mathbf{x_2}) = \cos(\mathbf{z}_{\mathbf{x_1}}, \mathbf{z}_{\mathbf{x_2}})/\tau. \tag{8}$$

The contrastive loss between image positive set $\{\mathbf{x}\}_+$ and image negative set $\{\mathbf{x}\}_-$ is

$$L_{intra_x} = \sum_{\mathbf{x_1} \in \{\mathbf{x}\}_\pm} -\frac{1}{n} \sum_{\mathbf{x_2} \in P(\mathbf{x_1})} \log \frac{\exp(\mathrm{S}_{\mathrm{inter}_{\mathrm{x}}}(\mathbf{x_1}, \mathbf{x_2}))}{\sum_{\mathbf{a} \in \{\mathbf{x}\}_\pm} \exp(\mathrm{S}_{\mathrm{inter}_{\mathrm{x}}}(\mathbf{x_1}, \mathbf{a}))}, \tag{9}$$

where $\{\mathbf{x}\}_\pm = \{\mathbf{x}\}_+ \cup \{\mathbf{x}\}_-$ and $P(\mathbf{x_1})$ is the set of images in the corresponding set $\{\mathbf{x}\}_+$ or $\{\mathbf{x}\}_-$ that differ from $\mathbf{x_1}$. Specifically, if $\mathbf{x_1}$ belongs to $\{\mathbf{x}\}_+$, then $P(\mathbf{x_1})$ corresponds to the collection of elements in $\{\mathbf{x}\}_+$ excluding $\mathbf{x_1}$. n represents the number of elements in set $P(\mathbf{x_1})$.

The computation for the music modality is similar to that for the image modality. Given music positive set $\{\mathbf{y}\}_+$ and music negative set $\{\mathbf{y}\}_-$, we define the score function between two pieces of music as

$$\mathrm{S}_{\mathrm{inter}_{\mathrm{y}}}(\mathbf{y_1}, \mathbf{y_2}) = \cos(\mathbf{z}_{\mathbf{y_1}}, \mathbf{z}_{\mathbf{y_2}})/\tau. \tag{10}$$

TABLE IV
ABLATION STUDY ON HYPER-PARAMETER n

Metric	GT	n=2	n=4	n=8	n=16
Polyphony Rate	0.5303	0.4713	0.4921	**0.5179**	0.5116
Pitch Entropy	3.9860	2.8129	3.2839	**3.5620**	3.3909
Groove Consistency	0.9922	0.8601	0.9229	**0.9760**	0.9578
Emotion Matching	0.1374	5.7670	2.9825	**1.8710**	3.2618

The best results are in boldface.

The contrastive loss between music positive set $\{\mathbf{y}\}_+$ and music negative set $\{\mathbf{y}\}_-$ is

$$L_{intra_y} = \sum_{\mathbf{y_1} \in \{\mathbf{y}\}_\pm} -\frac{1}{n} \sum_{\mathbf{y_2} \in P(\mathbf{y_1})} \log \frac{\exp(\mathrm{S}_{\mathrm{inter}_{\mathrm{y}}}(\mathbf{y_1}, \mathbf{y_2}))}{\sum_{\mathbf{a} \in \{\mathbf{y}\}_\pm} \exp(\mathrm{S}_{\mathrm{inter}_{\mathrm{y}}}(\mathbf{y_1}, \mathbf{a}))}, \tag{11}$$

where $\{\mathbf{y}\}_\pm = \{\mathbf{y}\}_+ \cup \{\mathbf{y}\}_-$ and $P(\mathbf{y_1})$ is the set of music in the corresponding set $\{\mathbf{y}\}_+$ or $\{\mathbf{y}\}_-$ that differ from $\mathbf{y_1}$. Specifically, if $\mathbf{y_1}$ belongs to $\{\mathbf{y}\}_+$, then $P(\mathbf{y_1})$ corresponds to the collection of elements in $\{\mathbf{y}\}_+$ excluding $\mathbf{y_1}$. n represents the number of elements in set $P(\mathbf{y_1})$.

To sum up, given images and music, we group image and music into positive set and negative set respectively, and the total loss during cross-modal contrastive learning is

$$L_c = L_{inter} + L_{intra_x} + L_{intra_y}. \tag{12}$$

Therefore, combined with the reconstruction loss (mentioned in (3)), the overall loss is defined as

$$L_{total} = L_{rec} + L_c. \tag{13}$$

V. EXPERIMENT

In this section, the experimental settings are introduced, including the implementation details and evaluation metrics. Subsequently, diverse image and music autoencoders are employed to validate the plug-and-play capability of the model. Then, we conduct comparison experiment with transformer and Generative Adversarial Network (GAN) to evaluate the experimental performance under the same amount of data. Finally, we conduct ablation experiments on the proposed intra-modal and inter-modal contrastive learning to verify the necessity of each contrastive learning component.

A. Experimental Settings

1) Implement Details: To validate the plug-and-play capability of the model, three sets of experiments are conducted using three image autoencoders and two music autoencoders to verify the effectiveness of the model. Specifically, we initially train the autoencoders on the proposed dataset. The image projector is composed of three fully connected layers with batch normalization (bn) layers and ReLU activation functions applied at each layer. Similarly, each of the music projector and de-projector consists of two fully connected layers, also incorporating bn layers and ReLU activation functions. The hyper-parameter value $n = 8$ is chosen for subsequent experiments based on the experimental results displayed in Table IV.

Authorized licensed use limited to: Hanyang University. Downloaded on December 20,2024 at 07:33:50 UTC from IEEE Xplore. Restrictions apply.

Image Autoencoders: Adversarial Latent AutoEncoders (ALAE) [44], Vector Quantized Variational Autoencoder (VQ-VAE) [45], and β-VAE [46] are adopted as image autoencoders respectively. ALAE is a type of generative model that combines the strengths of autoencoders and GANs to learn a disentangled representation of data through adversarial training. VQ-VAE is an autoencoder that learns a discrete codebook of embeddings through vector quantization for disentangled and interpretable data representations. β-VAE is a variant of Variational Autoencoder (VAE) that incorporates a regularization parameter to encourage disentangled and interpretable representations of data.

Music Autoencoders: Music FaderNets (FNT) [47] and Latent Space Regularization (LSR) [48] are adopted as music autoencoders. FNT uses Gaussian Mixture Variational Autoencoders (GM-VAEs) to perform semi-supervised clustering and infer high-level features from the low-level representations. LSR introduces a new method for regulating the feature space of a deep generative model by incorporating musically significant attributes along specific dimensions of the feature space to structure the feature space and make it more meaningful for music generation.

2) Evaluation Metrics: Multiple metrics are performed on the music generated based on the image and the evaluation metrics can be divided into two categories: those related to the quality of the generated music and those related to the emotion relevance to the original image.

Music Quality: Polyphony Rate [23] and *Pitch Entropy* [49] are adopted to evaluate the quality of notes in generated music, along with *Groove Consistency* [49] to evaluate the quality of rhythm in generated music. *Polyphony Rate* is a metric used to quantify the level of polyphony in a piece of music. It measures the degree to which multiple independent melodic lines or voices occur simultaneously within a musical composition. *Pitch Entropy* is a measure that quantifies the level of pitch variation or uncertainty in a musical composition. It assesses the diversity or distribution of pitches present in a piece of music. *Groove Consistency* is a metric used to evaluate the rhythmic stability and consistency in a music performance, particularly in the context of rhythmic patterns and timing. The formulas for these three metrics are presented as

$$\text{Polyphony Rate} = \frac{\text{MPA\_time\_steps}}{\text{time\_steps}},$$

$$\text{Pitch Entropy} = -\sum_{0}^{127} P(\text{pitch} = i) \log_2 P(\text{pitch} = i),$$

$$\text{Groove Consistency} = 1 - \frac{1}{T-1} \sum_{i=1}^{T-1} d(g_i, g_{i+1}), \tag{14}$$

where MPA_time_steps is the time steps where multiple piches are on, and time_steps is the total time steps. T is the number of measures, g_i is the binary onset vector of the i-th measure (a one at position that has an onset, otherwise a zero), and $d(g_i, g_{i+1})$ is the hamming distance between two vectors g and g'. Note that the quality of the generated music is determined by how close the metric is to the ground truth, rather than by the metric value being larger or smaller.

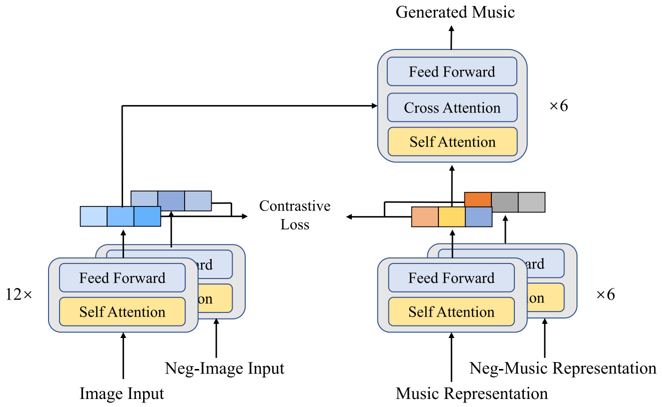

Fig. 4. Baseline: Transformer-based Image-to-Music Generation Model.

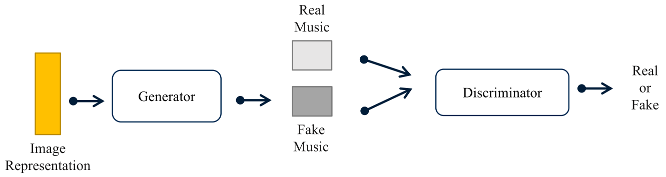

Fig. 5. Baseline: GAN-based Image-to-Music Generation Model.

Image-Music Emotion Alignment: To obtain the VA values of the generated music, a three-fully-connected-layer music VA predictor is trained to estimate the VA values of the generated music. The predicted VA values are then combined with those of the original image to calculate the emotion matching score between the image and the generated music. The average emotion matching score is referred to as the *Emotion Matching* metric.

3) Baselines: We compare our model with two simple baselines, Transformer-based Image-to-Music generation model (TF-IM) and GAN-based Image-to-Music generation model (GAN-IM).

TF-IM: The model consists of a 12-head image encoder, a 6-head music encoder, and a 6-head multi-modal encoder. Proposed contrastive learning are used to align two modalities. The model structure is shown as Fig. 4.

GAN-IM: Based on the GAN-based music generation network MidiNet [50], refinements and adaptations have been made to suit our task. MidiNet [50] generate melodies in the symbolic domain in a generative adversarial network (GAN) framework, along with a conditional mechanism to leverage prior knowledge and generate melodies with various constraints. Given the requirement to extract image features for music generation, adjustment is made to the generator. The revised generator comprises six downsampling layers and six upsampling layers, leveraging ResNet [51]. The discriminator in our model is structured similarly to MidiNet [50], comprising three convolution layers followed by one fully-connected layer. The model structure is shown as Fig. 5.

Authorized licensed use limited to: Hanyang University. Downloaded on December 20,2024 at 07:33:50 UTC from IEEE Xplore. Restrictions apply.

TABLE V
PERFORMANCE OF THE PROPOSED METHOD AND BASELINES

Metric	GT	ALAE+FNT	VQVAE+FNT	β-VAE+FNT	ALAE+LSR	VQVAE+LSR	β-VAE+LSR	TF-IM	GAN-IM
Polyphony Rate	0.5303	0.5156	**0.5179**	0.5053	0.5006	0.5019	0.4992	0.4104	0.3069
Pitch Entropy	3.9860	3.3632	**3.5620**	3.3231	3.2602	3.2803	3.2013	2.7635	1.8993
Groove Consistency	0.9922	**0.9777**	0.9760	0.9457	0.9241	0.9265	0.8928	0.8101	0.6132
Emotion Matching	0.1374	2.1327	**1.8710**	2.4609	2.7460	2.5149	2.9398	4.0601	4.9076

The best results are emphasized in bold. To validate the plug-and-play capability of the model, three image autoencoders (ALAE, VQVAE, and β-VAE) and two music autoencoders (FNT and LSR) are combined to form six sets of experiments. The baselines are transformer-based image-to-music generation model (denoted as TF-IM) and gan-based image-to-music generation model (denoted as GAN-IM).

TABLE VI
FURTHER EXPERIMENTS ON TF-IM

Metric	GT	TF/IMG	TF-FI	TF-IM
Polyphony Rate	0.5303	0.4983	**0.5094**	0.4104
Pitch Entropy	3.9860	**3.4126**	3.4022	2.7635
Groove Consistency	0.9922	**0.9389**	0.9110	0.8101

"TF/IMG" refers to TF-IM with its image branch removed, while "TF-FI" denotes TF-IM with fixed images as inputs.
The best results are in boldface.

B. Results

We combined ALAE, VQVAE, and β-VAE with FNT and LSR, resulting in six sets of autoencoder models as ALAE+FNT, VQVAE+FNT, β-VAE+FNT, ALAE+LSR, VQVAE+LSR, β-VAE+LSR. The comparison of the proposed method and baselines is shown in Table V. The following observations can be drawn from the results:

(1) In terms of *Emotion Matching*, the combination with VQ-VAE demonstrates superior performance compared to combining with ALAE or β-VAE. The reason for the superior performance of the VQVAE combined models in *Emotion Matching* can be attributed to the higher dimension of the feature space in VQVAE, which enables more distinct feature for different images and facilitates better modality alignment. This is also the reason why the ALAE combined models outperform the β-VAE combined models. In terms of *Music Quality*, using FNT as the music autoencoder performs better than LSR. This is reasonable because FNT encode a multi-dimensional, regularized feature space for each low-level feature, providing greater flexibility. With the plug-and-play capability, the proposed image-to-music generation model can integrate the advantages of existing autoencoders, resulting in efficient music generation.

(2) Trained on an equivalent amount of data, the proposed model outperforms TF-IM in terms of *Music Quality* and *Emotion Matching* metrics. To understand the reasons behind this performance gap, we further conduct experiments to investigate the underlying causes. We design two variants of TF-IM: excludes the image branch (denoted as TF/IMG) and takes a single fixed image as input (denoted as TF-FI). The results are as Table VI. It can be concluded that removing the image branch and using a single fixed image as input results in better performance in music generation. However, when the entire model is combined, the performance actually worsens. This is because adding high-dimensional image features requires the model to learn both music generation and modality alignment simultaneously, which interferes with the performance of a single music generation model, leading to inferior results.

(3) The proposed model outperforms GAN-IM. It is because music often has long-term dependencies, where notes and chords at different points in time are related to each other. GANs may struggle to capture these long-term dependencies, leading to generated music that sounds disjointed or lacks coherence. GANs are trained using a loss function that encourages the generator to produce outputs that are similar to the real data. This can sometimes lead to the generator producing outputs that are too similar to each other, resulting in a lack of diversity.

C. Ablation Studies

Different components of proposed contrastive learning and their impact is evaluated in this section. To study the effects of each contrastive learning component, four experiments are conducted: (1) inter-modal contrastive learning only, (2) intra-image and inter-modal contrastive learning, (3) intra-music and inter-modal contrastive learning, (4) intra-music, intra-image and inter-modal contrastive learning (5) no contrastive learning is employed as baseline. For baseline (5), We train a fully connected neural network with 5 layers to directly map the image feature vector to the music feature vector.

Contrastive learning: Table VII shows that using any of the contrastive learning component improves all metrics compared to the baseline. The largest improvement comes from the inter-modal contrastive learning and without contrastive learning, which improve *Emotion Matching* from 4.8211 to 3.2540. This suggests that the cross-modal latent space obtained through inter-modal contrastive learning is better than a direct mapping relationship between the image feature vector and the music feature vector. This is because inter-modal contrastive learning encourages the cross-modal latent space to capture meaningful semantic information that is shared across modalities, such as the emotion of an image and its corresponding music. In contrast, direct mapping may result in a loss of semantic information due to the limited capacity of the mapping function. Cross-modal latent space obtained through inter-modal contrastive learning is also more robust to noise and variation in the input data. Direct mapping may be more sensitive to noise and variation, as it relies solely on the input data and may not be able to capture the full range of relevant features.

Components of Contrastive Learning: Combining contrastive learning provides further gains. Experiments show that utilizing both intra-image and intra-music contrastive learning results in better performance than using either of them alone. This demonstrates that local and global conditions are complementary. The usage of intra-image and intra-music contrastive learning promotes the development of more resilient and distinct

Authorized licensed use limited to: Hanyang University. Downloaded on December 20,2024 at 07:33:50 UTC from IEEE Xplore. Restrictions apply.

TABLE VII
ABLATION STUDIES OF DIFFERENT COMPONENTS IN PROPOSED CONTRASTIVE LEARNING

Inter-Modal	Intra-Image	Intra-Music	Polyphony Rate	Pitch Entropy	Groove Consistency	Emotion Matching
✗	✗	✗	0.3882	3.0010	0.8351	4.8211
✓	✗	✗	0.4121	3.1403	0.8713	3.2540
✓	✓	✗	0.4643	3.3583	0.9264	2.8423
✓	✗	✓	0.4572	3.3992	0.9417	2.8964
✓	✓	✓	**0.5179**	**3.5620**	**0.9760**	**1.8710**

"INTER-MODAL", "INTRA-IMAGE" and "INTRA-MUSIC" denote the three different components of proposed contrastive learning. ✓means the corresponding component is utilized in the training process. For the experiment without contrastive learning, a fully connected neural network with five layers is trained to directly map the image feature vector to the music feature vector. The best results are in boldface.

latent vectors in the cross-modal latent space for each modality, which, in turn, helps mitigate overfitting to the training data by directing the model's attention towards the most pertinent modality-related features. As a result, combining inter-modal and intra-modal contrastive learning leads to better generalization performance.

VI. Conclusion

This paper presents a new task, i.e. generating a piece of pure music from a given natural image. Considering the ambiguity and subjectivity of the task, emotion is introduced to bridge the image and music on the human perception level. The proposed model maps the training images and music into two feature spaces respectively, which preserve the modality-related characteristics. Based on the learned feature spaces, a new contrastive learning strategy is designed to explore the shared emotion spaces, and establish the connection between the two modalities. In this way, the emotion consistent image-to-music translation can be achieved. Besides, a new cross-modal image-music dataset is established, which uses the MIDI format of music for a better representation, and employs continuous VA labels to annotate the emotions. Various metrics indicate that the music generated by our model exhibits high quality while maintaining consistency with the emotion of the corresponding images.

For future research, we intend to expand the scope of this approach to include other modalities. Although this paper primarily focuses on generating pure music from images, there is potential to apply this method to other modalities, such as generating background music for videos.

References

[1] H. L. Bonny, *Music and Consciousness: The Evolution of Guided Imagery and Music*. New Braunfels, TX, USA: Barcelona Publishers, 2002.

[2] X. Li, D. Hu, and X. Lu, "Image2song: Song retrieval via bridging image content and lyric words," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 5649–5658.

[3] Z. Xiong, P.-C. Lin, and A. Farjudian, "Retaining semantics in image to music conversion," in *Proc. IEEE Int. Symp. Multimedia*, 2022, pp. 228–235.

[4] X. Li, D. Tao, S. J. Maybank, and Y. Yuan, "Visual music and musical vision," *Neurocomputing*, vol. 71, no. 10, pp. 2023–2028, 2008.

[5] H.-T. Hung et al., "EMOPIA: A multi-modal pop piano dataset for emotion recognition and emotion-based music generation," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, 2021, pp. 318–325.

[6] A. Alajanki, Y.-H. Yang, and M. Soleymani, "Developing a benchmark for emotional analysis of music," *PLoS One*, vol. 12, no. 3, pp. 1–22, 2017.

[7] J. Fan, M. Thorogood, and P. Pasquier, "Emo-soundscapes: A. dataset for soundscape emotion recognition," in *Proc. IEEE Int. Conf. Affect. Comput. Intell. Interact.*, 2017, pp. 196–201.

[8] J. Fan, Y.-H. Yang, K. Dong, and P. Pasquier, "A comparative study of western and Chinese classical music based on soundscape models," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2020, pp. 521–525.

[9] J. Fan, K. Tatar, M. Thorogood, and P. Pasquier, "Ranking-based emotion recognition for experimental music," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, 2017, pp. 368–375.

[10] L. Ferreira and J. Whitehead, "Learning to generate music with sentiment," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, 2019, pp. 384–390.

[11] J. A. Mikels et al., "Emotional category data on images from the international affective picture system," *Behav. Res. Methods*, vol. 37, no. 4, 2005, Art. no. 626.

[12] P. Ekman, "An argument for basic emotions," *Cogn. Emotion*, vol. 6, no. 3-4, pp. 169–200, 1992.

[13] T. Xue, A. E. Ali, T. Zhang, G. Ding, and P. Cesar, "CEAP-360VR: A. continuous physiological and behavioral emotion annotation dataset for 360° VR videos," *IEEE Trans. Multimedia*, vol. 25, pp. 243–255, 2023.

[14] H.-R. Kim, Y.-S. Kim, S. J. Kim, and I.-K. Lee, "Building emotional machines: Recognizing image emotions through deep neural networks," *IEEE Trans. Multimedia*, vol. 20, pp. 2980–2992, 2018.

[15] S. Zhao et al., "Emotion-based end-to-end matching between image and music in valence-arousal space," in *Proc. ACM Int. Conf. Multimedia*, 2020, pp. 2945–2954.

[16] J.-C. Wang, Y.-H. Yang, H.-M. Wang, and S.-K. Jeng, "The acoustic emotion gaussians model for emotion-based music annotation and retrieval," in *Proc. ACM Int. Conf. Multimedia*, 2012, pp. 89–98.

[17] J.-C. Lin, W.-L. Wei, and H.-M. Wang, "EMV-matchmaker: Emotional temporal course modeling and matching for automatic music video generation," in *Proc. ACM Int. Conf. Multimedia*, 2015, pp. 899–902.

[18] J.-C. Lin, W.-L. Wei, and H.-M. Wang, "Automatic music video generation based on emotion-oriented pseudo song prediction and matching," in *Proc. ACM Int. Conf. Multimedia*, 2016, pp. 372–376.

[19] J.-C. Lin, W.-L. Wei, J. Yang, H.-M. Wang, and H.-Y. M. Liao, "Automatic music video generation based on simultaneous soundtrack recommendation and video editing," in *Proc. ACM Int. Conf. Multimedia*, 2017, pp. 519–527.

[20] Y. Wang, W. Liang, W. Li, D. Li, and L.-F. Yu, "Scene-aware background music synthesis," in *Proc. ACM Int. Conf. Multimedia*, 2020, pp. 1162–1170.

[21] G. Verma, E. G. Dhekane, and T. Guha, "Learning affective correspondence between music and image," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*,, 2019, pp. 3975–3979.

[22] X. Wu, Y. Qiao, X. Wang, and X. Tang, "Bridging music and image via cross-modal ranking analysis," *IEEE Trans. Multimedia*, vol. 18, pp. 1305–1318, 2016.

[23] H.-W. Dong, W.-Y. Hsiao, L.-C. Yang, and Y.-H. Yang, "MuseGAN: Multi-track sequential generative adversarial networks for symbolic music generation and accompaniment," in *Proc. AAAI Conf. Artif. Intell.*, 2018, pp. 34–41.

[24] Y.-N. Hung, I.-T. Chiang, Y.-A. Chen, and Y.-H. Yang, "Musical composition style transfer via disentangled timbre representations," in *Proc. Int. Joint Conf. Artif. Intell.*, 2019, pp. 4697–4703.

Authorized licensed use limited to: Hanyang University. Downloaded on December 20,2024 at 07:33:50 UTC from IEEE Xplore. Restrictions apply.

[25] C.-Y. Lu, M.-X. Xue, C.-C. Chang, C.-R. Lee, and L. Su, "Play as you like: Timbre-enhanced multi-modal music style transfer," in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 1061–1068.

[26] A. Roberts, J. Engel, C. Raffel, C. Hawthorne, and D. Eck, "A hierarchical latent vector model for learning long-term structure in music," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 4364–4373.

[27] O. Cıfka, A. Ozerov, U. Şimşekli, and G. Richard, "Self-supervised VQ-VAE for one-shot music style transfer," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2021, pp. 96–100.

[28] K. Tanaka, R. Nishikimi, Y. Bando, K. Yoshii, and S. Morishima, "Pitch-timbre disentanglement of musical instrument sounds based on VAE-based metric learning," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2021, pp. 111–115.

[29] Z. Hu, Y. Liu, G. Chen, and Y. Liu, "Can machines generate personalized music? A. hybrid favorite-aware method for user preference music transfer," *IEEE Trans. Multimedia*, vol. 25, pp. 2296–2308, 2023.

[30] Y.-S. Huang and Y.-H. Yang, "Pop music transformer: Beat-based modeling and generation of expressive pop piano compositions," in *Proc. ACM Int. Conf. Multimedia*, 2020, pp. 1180–1188.

[31] D. von Rtte, L. Biggio, Y. Kilcher, and T. Hofmann, "FIGARO: Controllable music generation using learned and expert features," in *Proc. Int. Conf. Learn. Representations*, 2023. [Online]. Available: https://openreview.net/forum?id=NyR8OZFHw6i

[32] X. Zhang, J. Zhang, Y. Qiu, L. Wang, and J. Zhou, "Structure-enhanced pop music generation via harmony-aware learning," in *Proc. ACM Int. Conf. Multimedia*, 2022, pp. 1204–1213.

[33] W. Ping, K. Peng, K. Zhao, and Z. Song, "Waveflow: A. compact flow-based model for raw audio," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 7706–7716.

[34] A. V. D. Oord et al., "Wavenet: A generative model for raw audio," 2016, *arXiv:1609.03499*.

[35] A. Santos, H. S. Pinto, R. P. Jorge, and N. Correia, "Music synthesis from images," in *Proc. Int. Conf. Comput. Creativity*, 2021, pp. 103–112.

[36] Y. Saito, H. Fujii, and S. Sagayama, "Semi-automatic music piece creation based on impression words extracted from object and background in color image," in *Proc. IEEE 10th Glob. Conf. Consum. Electron.*, 2021, pp. 268–272.

[37] S. Di et al., "Video background music generation with controllable music transformer," in *Proc. ACM Int. Conf. Multimedia*, 2021, pp. 2037–2045.

[38] X. Yang, Y. Yu, and X. Wu, "Double linear transformer for background music generation from videos," *Appl. Sci.*, vol. 12, no. 10, 2022, Art. no. 5050.

[39] L. Zhuo et al., "Video background music generation: Dataset, method and evaluation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2023, pp. 15637–15647.

[40] P. J. Lang, M. M. Bradley, and B. N. Cuthbert, "International affective picture system (IAPS): Technical manual and affective ratings," *NIMH Center Study Emotion Attention*, vol. 1, no. 39–58, p. 39, 1997.

[41] A. Marchewka, Ł. Żurawski, K. Jednoróg, and A. Grabowska, "The nencki affective picture system (NAPS): Introduction to a novel, standardized, wide-range, high-quality, realistic picture database," *Behav. Res. Methods*, vol. 46, pp. 596–610, 2014.

[42] S. R. Livingstone, R. Muhlberger, A. R. Brown, and W. F. Thompson, "Changing musical emotion: A. computational rule system for modifying score and performance," *Comput. Music J.*, vol. 34, no. 1, pp. 41–64, 2010.

[43] P. Khosla et al., "Supervised contrastive learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 18661–18673.

[44] S. Pidhorskyi, D. A. Adjeroh, and G. Doretto, "Adversarial latent autoencoders," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 14104–14113.

[45] A. van den Oord, O. Vinyals, and K. Kavukcuoglu, "Neural discrete representation learning," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 6309–6318.

[46] I. Higgins et al., "beta-VAE: Learning basic visual concepts with a constrained variational framework," in *Proc. Int. Conf. Learn. Representations*, 2017.

[47] H. H. Tan and D. Herremans, "Music FaderNets: Controllable music generation based on high-level features via low-level feature modelling," in *Proc. Int. Soc. Music Inf.*, 2020, pp. 109–116.

[48] A. Pati and A. Lerch, "Latent space regularization for explicit cntrol of musical attributes," in *Proc. Int. Conf. Mach. Learn. Mach. Learn. Music Discov. Workshop Mach. Learn. Music Discov. Workshop*, 2019.

[49] S. Wu and Y. Yang, "The jazz transformer on the front line: Exploring the shortcomings of AI-composed music through quantitative measures," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, 2020, pp. 142–149.

[50] L. Yang, S. Chou, and Y. Yang, "MidiNet: A convolutional generative adversarial network for symbolic-domain music generation," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, 2017, pp. 324–331.

[51] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

Yajie Wang is currently working toward the M.S. degree with the School of Artificial Intelligence, OPtics and ElectroNics (iOPEN), Northwestern Polytechnical University, Xi'an, China. Her research interests include machine learning and multi-modal retrieval.

Mulin Chen received the B.E. degree in software engineering and the Ph.D. degree in computer application technology from Northwestern Polytechnical University, Xi'an, China, in 2014 and 2019, respectively. He is currently an Associate Professor with the School of Artificial Intelligence, OPtics and ElectroNics (iOPEN), Northwestern Polytechnical University.

Xuelong Li (Fellow, IEEE) is currently a Full Professor with the School of Artificial Intelligence, OPtics and ElectroNics (iOPEN), Northwestern Polytechnical University, Xi'an, China.

Authorized licensed use limited to: Hanyang University. Downloaded on December 20,2024 at 07:33:50 UTC from IEEE Xplore. Restrictions apply.