Contrastive Audio-MIDI Learning for Symbolic-domain Musical Instrument Classification

1st Shun Sawada
Department of Information Technology and Media Design
Nippon Institute of Technology
Saitama, Japan
sawada.shun@nit.ac.jp

***Abstract*—Musical instrument classification in the symbolic domain is a challenging task due to the inherent differences between MIDI representations and their corresponding acoustic signals. In this study, we propose a contrastive learning-based framework, Contrastive Audio-MIDI Learning (CAMIL), to improve MIDI-domain instrument classification by leveraging both symbolic and audio information. In our approach, we use MIDI embeddings as anchors, pairing them with their corresponding audio embeddings as positive samples and audio embeddings from different instruments as negative samples. By optimizing a contrastive loss function, our model learns to align MIDI embeddings with their corresponding audio representations while pushing apart embeddings of different instruments. We evaluate our method on the Lakh MIDI dataset and demonstrate that it improves instrument classification performance in the symbolic domain. Our results highlight the potential of contrastive learning in bridging the gap between audio and MIDI representations for more robust musical instrument recognition.**

***Index Terms*—Musical instrument classification, Symbolic-domain, MIDI, Contrastive Learning**

I. Introduction

Musical instrument classification is a fundamental task in music information retrieval (MIR), supporting downstream applications such as automatic transcription, orchestration analysis, and composition assistance. Although instrument classification has been studied using symbolic (e.g., MIDI) and audio (e.g., spectrogram) representations, a substantial gap between these modalities hinders their integration for accurate and robust classification.

Previous work has primarily focused on approaches using audio signals. These classifications include methods targeting single notes, performance data of a certain length, and entire music pieces. Such models mainly learn and classify instruments based on their timbral features.

Although musical instrument classification using symbolic representations such as note sequences has been explored [1], [2], it has received less attention than its audio-based counterpart. This is partly due to the lack of large-scale symbolic datasets with reliable instrument labels and the inherent difficulty of inferring instruments from symbolic input, where timbral cues are not explicitly available.

Despite its difficulty, symbolic instrument classification is crucial for symbolic transcription, composition support, and semantic retrieval in music archives. Moreover, it provides useful representations for downstream tasks such as genre recognition, style transfer, and conditional music generation. Despite its challenges, this task remains highly valuable.

To bridge this gap, we propose Contrastive Audio-MIDI Learning (CAMIL), a novel framework that uses contrastive learning to align MIDI and audio representations of musical instruments. Each MIDI representation is treated as an anchor, paired with a positive audio sample from the same instrument and negative samples from different instruments. The model is trained to minimize the distance between matching Audio-MIDI pairs while maximizing the distance to mismatched ones, thereby learning a shared embedding space where semantically similar pairs are close and dissimilar ones are separated.

This study aims to enhance symbolic music processing by leveraging timbral information contained in audio signals. When audio is converted into symbolic representations such as musical scores, timbral characteristics are typically lost. However, audio signals capture not only pitch and duration but also rich information about timbre and expressive performance nuances. Instruments with similar physical structures often share timbral and performance characteristics. Appropriately extracting and associating this knowledge from audio with symbolic representations provides valuable cues for effective instrument classification.

Contrastive learning has proven effective in aligning representations across different modalities. CLAP (Contrastive Language-Audio Pretraining) [3] learns joint embeddings for audio and text, while CLIP [4] does the same for vision-language tasks. Inspired by these methods, we apply contrastive learning to connect MIDI and audio representations, enabling robust instrument classification.

The contributions of this paper are as follows:

- We propose Contrastive Audio-MIDI Learning (CAMIL), a novel framework that aligns MIDI and audio representations via contrastive learning for robust symbolic-domain instrument classification.
- We demonstrate that CAMIL achieves comparable or superior performance to supervised baselines, even under unsupervised settings.
- We evaluate CAMIL on the Lakh MIDI Dataset, analyzing data efficiency and the effect of different negative

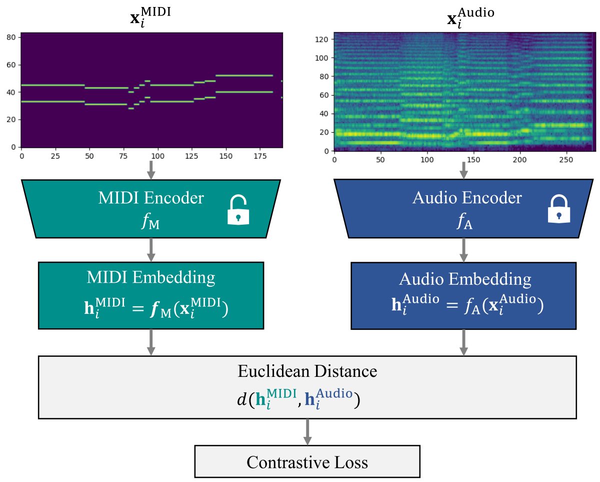
$\mathbf{x}_i^{\text{MIDI}}$
$\mathbf{x}_i^{\text{Audio}}$
80
60
40
20
0
0
25
50
75
100
125
150
175
120
100
80
60
40
20
0
0
50
100
150
200
250
MIDI Encoder
f_{M}
Audio Encoder
f_{A}
MIDI Embedding
$\mathbf{h}_i^{\text{MIDI}} = f_{\text{M}}(\mathbf{x}_i^{\text{MIDI}})$
Audio Embedding
$\mathbf{h}_i^{\text{Audio}} = f_{\text{A}}(\mathbf{x}_i^{\text{Audio}})$
Euclidean Distance
$d(\mathbf{h}_i^{\text{MIDI}}, \mathbf{h}_i^{\text{Audio}})$
Contrastive Loss

Fig. 1: Network Architecture of the Contrastive Audio-MIDI Learning

sampling strategies.

II. Related Work

Research on musical instrument classification has predominantly focused on audio-based methods. Various techniques, including machine learning and deep learning approaches, have been used to classify musical instruments based on their timbral features. Recent works have leveraged deep learning architectures, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), for instrument classification tasks [5], [6]. In addition, hierarchical classification methods have been proposed to improve generalization to unseen instrument classes by exploiting inter-instrument relationships. For example, Garcia et al. [7] introduce a hierarchical few-shot learning approach that leverages instrument taxonomy to improve accuracy with limited data.

Compared to audio-based classification, instrument classification from symbolic representations has been less explored. Ji et al. proposed a language model-based approach to classify eight monophonic instruments from MIDI sequences [1]. Sawada proposed a deep learning-based framework that utilizes MIDI note sequences for instrument classification. The method improves instrument classification by distilling knowledge from an audio signal model (teacher model) into a MIDI sequence model (student model) [2].

Previous studies have explored cross-modal retrieval by learning correspondences between sheet music images and audio [8], [9], primarily to improve music retrieval. In contrast, our goal is to improve symbolic-domain instrument classification by leveraging timbral information from audio signals through contrastive learning. We propose to align MIDI and audio representations during training to enrich symbolic representations with acoustic characteristics, which are typically lost when converting audio into symbolic formats such as scores.

Multimodal learning has been widely investigated in various MIR tasks, including music emotion classification and music generation [10], [11]. Deep learning methods have been employed to model relationships between different modalities, such as MIDI and audio, to improve classification performance [12]. However, many existing approaches require both modalities during inference, which limits their practical application. Our method differs in that we leverage audio information only during training to enhance MIDI-based classification, ensuring that no audio data is needed at inference time.

Contrastive learning has proven effective in cross-modal representation learning. For example, CLIP [4] and CLAP [3] align vision-language and audio-language modalities, respectively, by learning joint embeddings. Inspired by these approaches, we apply contrastive learning to bridge MIDI and audio representations, enabling symbolic-domain instrument classification without requiring audio input at inference time.

III. Symbolic-domain Musical Instrument Classification

A. Musical Instrument Classification Task Setting

Musical instrument classification has been explored across diverse input types and task settings. In the audio domain, targets include single notes, time-series segments, or entire tracks, while in the symbolic domain, classification is typically performed on monophonic or polyphonic note sequences. Depending on the setting, the task can be formulated as either single-label or multi-label classification. Our work focuses on symbolic-domain classification using phrase-level polyphonic sequences, where timbral cues are not directly available.

In this study, we address symbolic-domain instrument classification from note sequences, following the setting of [2]. Given a segment of MIDI data, our goal is to predict the performing instrument based solely on its note sequence, regardless of whether it is monophonic or polyphonic. We represent MIDI data as piano rolls and focus on both temporal transition patterns and simultaneous note structures. The classification is performed on two-measure segments.

Let $X = \{x_1, x_2, \ldots, x_N\}$ be a set of symbolic music segments extracted from MIDI sequences. In this study, we focus on classifying musical instruments from note sequences in the symbolic domain. Each segment is extracted as a fixed-length span corresponding to two measures within a MIDI sequence.

The goal of the task is to predict the instrument labels associated with a given note sequence segment x. We define the set of instrument labels as $Y = \{1, 2, \ldots, K\}$, where K is the total number of instrument classes, and $y \in Y$ denotes the ground truth label for the segment. We formulate the instrument classification task as a multi-class classification problem, where the model f maps a note sequence segment x to a single instrument label, such that $f : x \mapsto y$. Each segment is treated as an independent sample during training and inference, and the model predicts the instrument label for each segment individually.

B. Contrastive Audio-MIDI Learning (CAMIL) for Symbolic-domain

We propose Contrastive Audio-MIDI Learning (CAMIL) to bridge the gap between MIDI-based and audio-based representations of musical instruments (see **Fig. 1**). Our model learns a joint embedding space for symbolic and audio representations of music segments. Given an input symbolic sequence and its corresponding audio representation, we employ two separate encoders: f_{M} for MIDI-based melodies and f_{A} for audio. These encoders transform the inputs into a common feature space as follows: $\mathbf{h}_{\mathrm{MIDI}} = f_{\mathrm{M}}(\mathbf{x}_{\mathrm{MIDI}}), \quad \mathbf{h}_{\mathrm{Audio}} = f_{\mathrm{A}}(\mathbf{x}_{\mathrm{Audio}})$, where $\mathbf{x}_{\mathrm{MIDI}}$ and $\mathbf{x}_{\mathrm{Audio}}$ denote the MIDI and audio inputs, respectively, and $\mathbf{h}_{\mathrm{MIDI}}, \mathbf{h}_{\mathrm{Audio}}$ are their corresponding embeddings.

To measure the similarity between embeddings, we compute the Euclidean distance $d = \|\mathbf{h}_{\mathrm{MIDI}} - \mathbf{h}_{\mathrm{Audio}}\|_2$. The model is trained to minimize this distance d for positive pairs (i.e., matching Audio-MIDI pairs), while ensuring that negative pairs (non-matching pairs) are pushed apart using a contrastive loss function (See Section III-D).

C. Positive and Negative Pair Selection

In our contrastive learning framework, we define the anchor and its corresponding positive and negative samples as follows:

- **Anchor**: A MIDI-based representation of a musical segment played by a specific instrument.
- **Positive sample**: The corresponding audio representation of the same musical segment, aligned with the anchor.
- **Negative sample**: An audio representation from a different instrument or from a randomly selected segment that is not aligned with the anchor.

Given a dataset containing N samples, each MIDI representation $\mathbf{x}_i^{\mathrm{MIDI}}$ is paired with a corresponding audio representation $\mathbf{x}_i^{\mathrm{Audio}}$, forming a positive pair:

$$P_{\text{positive}} = \{(\mathbf{x}_i^{\mathrm{MIDI}}, \mathbf{x}_i^{\mathrm{Audio}}) \mid i = 1, 2, \ldots, N\}. \quad (1)$$

In our framework, we consider two types of negative sampling strategies:

- **Random negatives (unsupervised setting)**: Negative samples are randomly selected from the dataset without considering class information. As a result, negative pairs are not semantically aligned with the anchor and may still belong to the same instrument class. Formally, random negatives are defined as:

$$P_{\text{negative}} = \{(\mathbf{x}_i^{\mathrm{MIDI}}, \mathbf{x}_j^{\mathrm{Audio}}) \mid i \neq j\}. \quad (2)$$

- **Class-based negatives (supervised setting)**: Negative samples are selected from audio representations belonging to different instrument classes. This strategy leverages class labels to ensure stronger semantic separation. The class-based negative pairs are defined as:

$$P_{\text{negative}} = \{(\mathbf{x}_i^{\mathrm{MIDI}}, \mathbf{x}_j^{\mathrm{Audio}}) \mid c_i \neq c_j\}, \quad (3)$$

where c_i and c_j denote the instrument class labels of the i-th and j-th samples, respectively.

By incorporating both random and class-based negative sampling strategies, our framework supports both supervised and unsupervised training settings.

D. Loss Function

In order to learn meaningful representations of the input data, we adopt a contrastive loss [13] during the pre-training phase. This loss encourages the model to minimize the distance between matching pairs (positive pairs) and maximize the distance between non-matching pairs (negative pairs), helping the model to distinguish between similar and dissimilar inputs. The contrastive loss $\mathcal{L}_{\text{pre}}$ is defined as follows:

$$\mathcal{L}_{\text{pre}} = \mathbb{E}\left[z \cdot d^2 + (1 - z) \cdot \max(0, \alpha - d)^2\right], \quad (4)$$

where $z \in \{0, 1\}$ is a binary label indicating whether the pair is a positive match ($z = 1$) or a negative pair ($z = 0$), and α is a margin hyperparameter that separates negative pairs beyond a certain distance. d represents the distance between embeddings.

Once the model has been pre-trained, we proceed to fine-tune it for the specific task of instrument classification. For this fine-tuning step, we use a cross-entropy loss to optimize the model.

E. Model Configuration

As shown in **Fig. 1**, our model consists of two encoders: a **MIDI encoder** and an **audio encoder**. Both encoders share a common architecture based on ResNet [14], which is known for its ability to learn deep representations through residual connections. The ResNet architecture has been successful in various domains, making it an ideal choice for our task of extracting meaningful features from both MIDI and audio data.

Both encoders are based on a ResNet architecture. The **Initial Convolution and Pooling** stage consists of a 7 × 7 convolutional layer with 64 filters and a stride of 2, followed by batch normalization and ReLU activation. A 3 × 3 max pooling layer with a stride of 2 is then applied. The **Residual Blocks** are organized into four stages with filter sizes of 64, 128, 256, and 512. Downsampling is applied at the first block of each stage using a stride of 2. Finally, a **Global Average Pooling Layer** reduces the spatial dimensions, and a **Fully Connected Layer** outputs the feature embedding.

IV. Evaluation

A. Datasets

While there are numerous datasets for instrument classification targeting audio signals, many of them focus on single note instrument sounds. To the best of our knowledge, there is no existing open datasets for instrument classification targeting note sequences. There are datasets consisting of pairs of audio signals from instrument performances and musical note sequences for automatic music transcription purposes. To acquire paired data of musical note sequences and audio signals for evaluation purposes, the Synthesized Lakh (**Slakh**) Dataset is utilized, which is a dataset for audio source separation that is synthesized from the **Lakh MIDI** Dataset [15]. The ground truth data for instrument classification consisted of the following eight categories selected from the Slakh Dataset: Guitar, Piano, Bass, Strings, Organ, Brass, Pipe, and Reed.

B. Pre-processing

We use the Lakh MIDI Dataset as our dataset for symbolic music. Each MIDI file contains multiple instrument tracks, which are paired with corresponding audio recordings to create Audio-MIDI training pairs.

MIDI preprocessing follows the method described in prior work [2]. First, non-melodic and percussion tracks are filtered out to retain only pitched instrument tracks. Then, velocity values are normalized to ensure consistency across MIDI files. Finally, each MIDI sequence is converted into a fixed-length piano-roll representation. We set the temporal resolution such that each measure is divided into 24 time steps, allowing the representation of common rhythmic patterns such as triplets and 32nd notes. Each track is segmented into two-measure units. Based on an analysis of pitch distributions, we limit the pitch range to 84 distinct values, spanning from C1 to B7. As a result, a 4/4 bar with a single track is represented as a 96 × 84 matrix, and each two-measure segment becomes a tensor of size 16,128 = 84 (pitches) × 192 (time steps).

For audio processing, we extract mel-spectrograms from paired audio recordings. The spectrogram representations are aligned with MIDI sequences by dividing them into segments corresponding to measure lengths, based on the BPM (Beats Per Minute) of the piece. The window size and hop size for the Fourier transform were set to 512, and the number of bins in the mel-filterbank was 128. The resulting size of the target output tensor was 36,096 = 128 (bins) × 282 (time steps).

C. Experimental Setup

We compare five settings to evaluate the effectiveness of contrastive Audio-MIDI learning:

- **AIC (Audio encoder)**: A classifier is trained on audio signals to obtain embeddings, which are used as reference representations in contrastive Audio-MIDI learning.
- **SIC (Baseline)**: A supervised classifier is trained directly on MIDI embeddings without contrastive Audio-MIDI learning.
- **CAMIL**: MIDI embeddings are pre-trained using contrastive Audio-MIDI learning with class-based negative sampling (see Eq. 3). The encoder is frozen, and a linear classifier is trained on top.
- **Fine-tuned CAMIL**: Same as CAMIL, but the encoder is fine-tuned during classifier training.
- **Unsupervised CAMIL**: MIDI embeddings are pre-trained using contrastive Audio-MIDI learning with random negative sampling (see Eq. 2). The encoder is then frozen, and a linear classifier is trained.

TABLE I: Classification performance (F1 Score) under different experimental conditions.

Condition	**Train data**	**Negative Sampling**	**F1 Score**
AIC (Audio encoder)	100%	-	0.9613
SIC (Baseline)	100%	-	0.6008
CAMIL	100%	Class-based	0.6236
Fine-tuned CAMIL	100%	Class-based	0.6270
Unsupervised CAMIL	100%	Random	0.6086
SIC (Baseline)	50%	-	0.5573
CAMIL	50%	Class-based	0.5772
Unsupervised CAMIL	50%	Random	0.5768

To evaluate data efficiency, we conduct experiments using both 100% and 50% of the training data. In all settings, the test set is used in full.

The model was trained using a batch size of 512 and an initial learning rate of 0.001. The Adam optimizer was employed for parameter optimization.

D. Results and Discussion

Tab. I summarizes the classification performance (F1 Score) under different experimental conditions. The results highlight the impact of contrastive learning and different negative sampling strategies on classification accuracy. First, the AIC (Audio encoder) model, serving as a reference performance, achieves an F1 score of 0.9613. This could serve as an upper bound for the performance of MIDI-based classification models.

The SIC (Baseline) model, which does not incorporate contrastive learning, exhibits lower performance with an F1 score of 0.6008 when trained on the full dataset. This suggests that direct classification using MIDI embeddings without pre-training struggles to capture discriminative features effectively.

Introducing CAMIL (Contrastive Audio-MIDI Learning) improves the performance compared to the SIC baseline. CAMIL with class-based negative sampling achieves an F1 score of 0.6236, and fine-tuning further enhances it to 0.6270. The results indicate that pre-training MIDI embeddings using contrastive learning effectively enhances feature representations, leading to improved classification accuracy.

When employing random negative sampling in an unsupervised setting, CAMIL achieves an F1 score of 0.6086. This is lower than the class-based approach, suggesting that selecting negatives from different instrument classes provides more effective supervision compared to purely random negatives. Notably, the unsupervised CAMIL model, which does not rely on instrument labels for negative sampling, slightly outperforming SIC. This suggests that contrastive learning can provide benefits even without explicit instrument classification labels, making it a promising approach for applications where labeled data is limited.

Reducing the training dataset to 50% leads to performance degradation across all models. The SIC baseline drops to 0.5573, highlighting the importance of sufficient training data. However, CAMIL still outperforms the SIC baseline, with

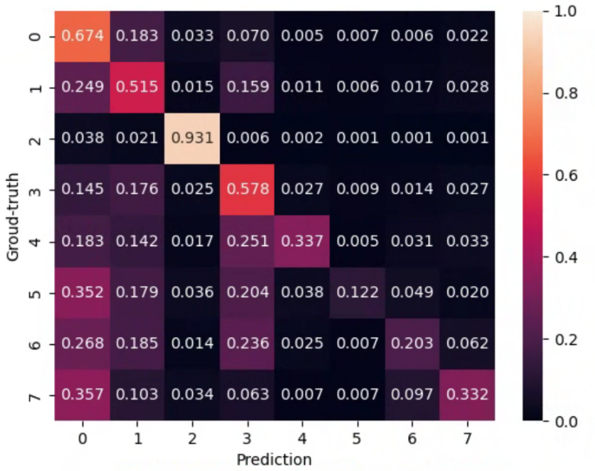

Fig. 2: Confusion matrix for the CAMIL, showing the classification results for the following instrument categories: Guitar (0), Piano (1), Bass (2), Strings (3), Organ (4), Brass (5), Pipe (6), and Reed (7).

the class-based negative sampling approach reaching 0.5772 and the random negative sampling method achieving 0.5768. These results indicate that CAMIL provides more robust representations that maintain relative performance advantages even with limited training data.

To better understand the classification behavior of our model, we present the confusion matrix of CAMIL in **Fig. 2**. It illustrates classification accuracy across instrument classes and highlights common misclassification patterns. CAMIL achieves high accuracy for well-represented classes such as "Piano" (class 1) and "Strings" (class 3). while relatively rare classes like "Flute" (class 6) and "Oboe" (class 7) exhibit higher misclassification rates. This suggests that CAMIL, despite its overall effectiveness, may be sensitive to class imbalance. These observations indicate that CAMIL learns robust feature representations for frequent classes, and emphasize the importance of class distribution in training effective classifiers.

V. Conclusion

In this study, we proposed Contrastive Audio-MIDI Learning (CAMIL) to improve symbolic-domain musical instrument classification by bridging the gap between MIDI representations and their corresponding audio features. Our method leverages contrastive learning by aligning MIDI embeddings with audio representations from the same segment, while pushing them away from unrelated samples, thereby enhancing representation consistency.

Experiments on the Lakh MIDI Dataset demonstrated that CAMIL improves classification performance, achieving an F1 score of 0.6270, outperforming the baseline (0.6008). Notably, even in an unsupervised setting, where instrument labels are not used for negative sampling, CAMIL achieved an F1 score of 0.6086, matching or slightly exceeding the baseline, indicating its effectiveness without relying on explicit class supervision. Furthermore, the model maintained robust performance even when the training data was reduced to 50%, suggesting that CAMIL is a data-efficient approach. These results indicate that contrastive learning with audio guidance enables the acquisition of meaningful MIDI representations, even in low-resource or weakly supervised scenarios.

For future work, we plan to analyze the generalizability of the learned representations and evaluate their transferability to other music-related tasks. This study highlights the potential of contrastive learning in symbolic music processing and opens new directions for building more robust and expressive MIDI-based models.

References

[1] K. Ji, D. Yang, and T. Tsai, "Instrument classification of solo sheet music images," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 546–550.

[2] S. Sawada, "Symbolic-domain musical instrument classification using knowledge distillation from audio-teacher to symbolic-student," in *2024 32nd European Signal Processing Conference (EUSIPCO)*. IEEE, 2024, pp. 191–195.

[3] Y. Wu, K. Chen, T. Zhang, Y. Hui, T. Berg-Kirkpatrick, and S. Dubnov, "Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[4] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PmLR, 2021, pp. 8748–8763.

[5] Z. Zhong, M. Hirano, K. Shimada, K. Tateishi, S. Takahashi, and Y. Mitsufuji, "An attention-based approach to hierarchical multi-label music instrument classification," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[6] M. Krause and M. Müller, "Hierarchical classification for instrument activity detection in orchestral music recordings," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023.

[7] H. F. Garcia, A. Aguilar, E. Manilow, and B. Pardo, "Leveraging hierarchical structures for few-shot musical instrument recognition," *arXiv preprint arXiv:2107.07029*, 2021.

[8] M. Dorfer, J. Hajič Jr, A. Arzt, H. Frostel, and G. Widmer, "Learning audio–sheet music correspondences for cross-modal retrieval and piece identification," *Transactions of the International Society for Music Information Retrieval*, vol. 1, no. 1, 2018.

[9] L. Carvalho, T. Washüttl, and G. Widmer, "Self-supervised contrastive learning for robust audio-sheet music retrieval systems," in *Proceedings of the 14th ACM Multimedia Systems Conference*, ser. MMSys '23. New York, NY, USA: Association for Computing Machinery, 2023, p. 239–248. [Online]. Available: https://doi.org/10.1145/3587819.3590968

[10] Q. Lu, X. Chen, D. Yang, and J. Wang, "Boosting for multi-modal music emotion," in *International Society for Music Information and Retrieval Conference*, 2010, pp. 105–105.

[11] D. Guan, X. Chen, and D. Yang, "Music emotion regression based on multi-modal features," in *International Symposium on Computer Music Modeling and Retrieval*, 2012, pp. 70–77.

[12] J. Zhou, X. Chen, and D. Yang, "Multimodel music emotion recognition using unsupervised deep neural networks," in *Conference on Sound and Music Technology (CSMT) Revised Selected Papers*, 2019, pp. 27–39.

[13] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06)*, vol. 2. IEEE, 2006, pp. 1735–1742.

[14] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[15] E. Manilow, G. Wichern, P. Seetharaman, and J. Le Roux, "Cutting music source separation some Slakh: A dataset to study the impact of training data quality and quantity," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2019.