INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

# Joint Time and Frequency Transformer for Chinese Opera Classification

*Qiang Li, Beibei Hu*

Ajmide Media, Shanghai, China
liqiang@ajmide.com, hubeibei@ajmide.com

## Abstract

Transformer has recently gained more attention and is widely used in audio tasks. Most tasks compute attention directly over the entire time-frequency space or only in the temporal. This paper presents a joint time and frequency model for Chinese opera classification. A shallow convolutional block is used to get localized low-level semantic features and reduce the feature map size. Moreover, the criss-cross attention and the factorised self-attention are employed in the model to extract the time and frequency space representation. The experiment results demonstrate that the proposed model achieves state-of-the-art performance on a large Chinese opera dataset with fewer model parameters.

**Index Terms**: time attention, frequency attention, Chinese opera classification, Transformer

## 1. Introduction

Chinese opera is a comprehensive stage art style with a long history in China. Moreover, how to better inherit and develop opera has always been an essential part of Chinese culture in every era. Typically, opera classification is used as a fundamental understanding of the field of opera. However, it has significant research and application value for downstream tasks, such as establishing a database of opera materials [1], analyzing the characteristics of the singing style and structure for Peking opera [2, 3], etc. [4, 5]

However, there are some challenges in Chinese opera classification. The first is the difficulty of data annotation for different opera genres because it can only be done by specific groups of people, such as opera performers or loyal listeners. Then, unlike music, the non-silence opera fragment is mainly composed of three parts: pure music, song, and speech [1]. So, it has a more complex content expression. Finally, during the long period of forming and developing traditional Chinese opera, the different operas influenced and cross-pollinated one another, resulting in remarkable similarities between different operas.

With the introduction of Vision Transformer (ViT) [6], a purely attention-based model which uses patch embedding to replace convolutional neural networks (CNNs) is widely applied for many tasks [7, 8, 9]. One disadvantage of self-attention in standard Transformer is the high computational complexity of the model that requires computing a similarity measure for all time-frequency bins. To increase computational efficiency and model performance, this paper proposes to combine Transformer with CNNs for the Chinese opera classification task.

First, a convolutional block is proposed to extract the latent representation of the Mel-spectrogram. Unlike patch embedding in the AST model [7], which needs to subdivide Mel-spectrogram evenly into patches, the convolutional block provided added flexibility.

Second, based on the attention mechanism, Transformer learns a representation by relating different positions in sequences. However, most of it is used as a temporal feature aggregator, which only computes the correlation between the temporal locations after combining frequency dimension with channel features, or space feature aggregator by computing the attention of the patch over the entire time-frequency space. However, it is well known that different audio components exist in different frequency ranges, and there is a robust spectral correlation. Therefore, this paper employs the criss-cross attention and the factorised self-attention to compute attention in a Transformer: consider both the horizontal and vertical dependencies of time-frequency bins simultaneously or separately in attention.

In summary, the contributions of the paper are as follows: (1) This paper introduces a novel joint time and frequency transformer for Chinese opera classification. (2) Experimentally, the proposed model achieved state-of-the-art results for Chinese opera classification.

## 2. Related work

To our knowledge, there is little research on Chinese opera classification tasks. In [10], seven machine-learning classifiers and ten hand-crafted acoustic features were applied and compared to test the classification results of eight typical genres of traditional Chinese opera. In [11], it used multi-feature fusion and extreme learning machines to discriminate eight typical genres. In [12], designed for music auto-tagging, the Musicnn [13] model was adopted to classify 18 Chinese opera genres containing horizontal and vertical convolutional filters.

With the advancement of deep learning, architectures based on CNN or Transformer have been continuously proposed and applied to different tasks. Like [12], we investigate some methods for audio classification tasks. Based on different CNN variants, [14] conducts a consistent evaluation of different music tagging models. In [15, 16], the sequence modeling approach adopted a Transformer to summarize the temporal sequence of the extracted local features by CNN. This method combines the frequency dimension and channel features for input to the Transformer encoder. [7] proposed audio spectrogram Transformers (AST) for audio classification tasks, which are purely based on self-attention. Due to the dependence on time frames and frequency bins, [17] proposes SpecTNT, a time-frequency transformer that models spectrograms as a sequence along both time- and frequency-axes. Similar to the temporal Transformer and spectral Transformer in [17], we also investigated different axis-attention methods for image or video tasks [8, 9, 18, 19], which there is no related method in audio applications. Those


10.21437/Interspeech.2023-1582

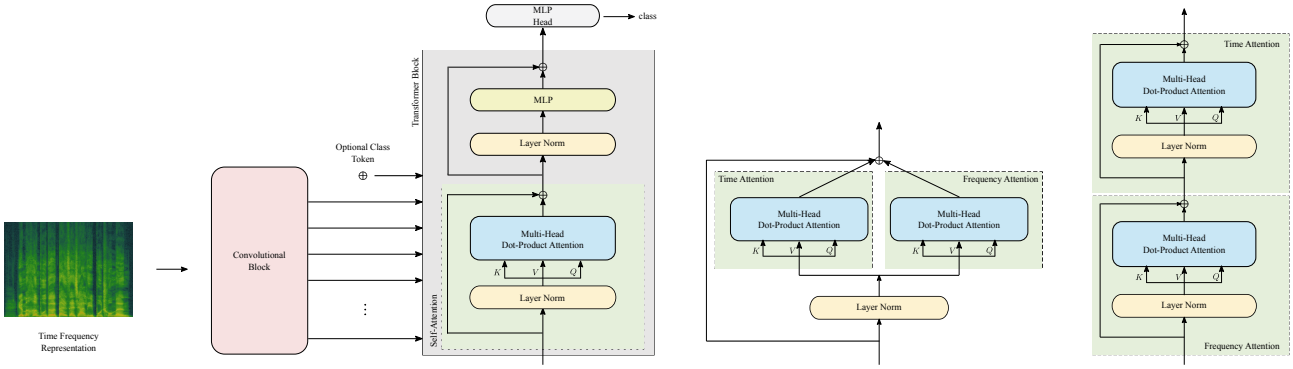


Figure 1: *Overall architecture of the proposed model shown on the left for the Chinese opera classification. Two attention patterns over time and frequency are shown on the right.*

methods mainly calculate the attention on a single axis, such as time, space for video or height, and width for image, and then combine them in different ways.

## 3. Methodology

As illustrated in Figure 1, we first apply a convolutional block for local feature aggregation. Then, the representation is fed into various Transformer models to capture long-distance feature dependencies and obtain embeddings. Finally, the embeddings are used for classification with a linear layer. The detailed architecture of the proposed network is discussed in the following subsections.

### 3.1. Convolutional block

In AST model, the input of Transformer is a patch embedding that is obtained through a linear projection layer after evenly subdividing a time-frequency representation into patches. To replace the above two operations and reduce feature map size, this paper proposes to use a convolutional block to capture low-level semantic features.

Instead of using a lightweight block with only one convolution layer or ResNet [20] model with complex deep convolution, we use shallow convolutional layers as our convolutional block. Like [14], we use $3 \times 3$ convolution filters with residual connections on Mel-spectrogram inputs. And instead of using seven layers, we only use a four-layered convolutional block. The representation after a convolutional block is denoted as $X \in R^{T \times F \times C}$, where $F$ is the number of mel bins, $T$ is the number of time steps, and $C$ is the number of attention channels of Transformer.

### 3.2. Transformer models

This section introduces the time and frequency attentions computed differently in Transformer. To start with, we introduce the scaled dot-product attention first. Then the criss-cross attention [18] and the factorised self-attention [9] will be explained.

#### *3.2.1. Scaled dot-product attention*

Scaled dot-product attention plays a pivotal role in the Multi-Head Self-Attention layer (MHSA) of Transformer [21]. MHSA first generates a set of queries $Q \in R^{N \times d}$, keys $K \in R^{N \times d}$, values $V \in R^{N \times d}$ with the corresponding projection. Then the query vector $q \in R^d$ is matched against each key vector in $K$. The output is the weighted sum of a set of $N$ value vectors $v$ based on the matching score. This process is called scaled dot-product attention:

$$Attention(Q, K, V) = Softmax\left(\frac{QK^T}{\sqrt{d}}\right)V \quad (1)$$

To prevent minimal gradients and stabilize the training process, each element in $QK^T$ multiplies by a constant $1/\sqrt{d}$ to be rescaled into a standard range.

#### *3.2.2. Criss-cross attention in both time and frequency directions*

Based on speech signal processing, different frequency bins have specific dependencies, and the time frames also have the same property. Instead of the entire feature space, criss-cross attention [18] considers feature points located in the time and frequency directions when computing the MHSA. This method can reduce the size of the $N$ value representing computational complexity, and experimental results show that it also leads to improved classification accuracy.

As shown in the middle of Figure 1, given local feature maps $X \in R^{T \times F \times C}$, the module firstly applies three convolutional layers with $1 \times 1$ filters on $X$ to generate three feature maps $Q$, $K$ and $V$, respectively, where $\{Q, K, V\} \in R^{T \times F \times C'}$. To get the attention of each feature point, compute it respectively in the time and frequency directions. In the Frequency direction, we can obtain the similarity matrix $D_F \in R^{T \times F \times F}$ by directly computing the $QK^T$ operation. In the time direction, we can also obtain the similarity matrix $D_T \in R^{F \times T \times T}$ through the similar processing steps above. After reshaping the matrix $D_T$ from $R^{F \times T \times T}$ to $R^{T \times F \times T}$, we apply concatenation to the two similarities matrix to get $D \in R^{T \times F \times (T+F)}$. Then, a softmax layer is applied on $D$ over the last dimension to calculate map $A \in R^{T \times F \times (T+F)}$. Finally, to obtain attention, the following processing steps are applied:

$$X' = reshape(A_T V_T) + A_F V_F + X \quad (2)$$

where $A_T \in R^{F \times T \times T}$ and $A_F \in R^{T \times F \times F}$ is part of $A$. $V_T \in R^{F \times T \times C'}$ and $V_F \in R^{T \times F \times C'}$ are equal to $V$. $X' \in R^{T \times F \times C'}$. $reshape$ is used to transpose the time and frequency dimension.

*3.2.3. Factorised self-attention in both time and frequency directions*

Different from criss-cross attention, factorised self-attention [9] computes time and frequency sequentially, as shown in the right of Figure 1.

After generating three feature maps $\{Q, K, V\} \in R^{T\times F\times C'}$, the frequency attention $X'_F \in R^{T\times F\times C'}$ can be obtained using Equation 3 which introduces residual operations. Like Equation 3, the time attention is computed in Equation 4. The input is $X''_F$ by reshaping $X'_F$ from $R^{T\times F\times C'}$ to $R^{F\times T\times C'}$. Moreover, three feature maps $\{Q_T, K_T, V_T\} \in R^{F\times T\times C'}$ can be obtained by applying three convolutional layers on $X''_F$. Finally, attention $X' \in R^{F\times T\times C'}$ can be obtained.

$$X'_F = Softmax\left(\frac{QK^T}{\sqrt{C'}}\right)V + X \tag{3}$$

$$X' = Softmax\left(\frac{Q_T K_T^T}{\sqrt{C'}}\right)V_T + X''_F \tag{4}$$

# 4. Dataset

This section gives some introduction to our Chinese opera dataset. According to the statistics, about 360 kinds of opera genres in various regions of China contain a great deal of complexity and variety [22]. We selected 21 genres as the classification objects based on the regional location and singing style [10]. They are: Teochew Opera (潮剧), Kunqu Opera (昆曲), Sichuan Opera (川剧), Hebei Clapper Opera (河北梆子), Shanghai Opera (沪剧), Huagu Opera (花鼓戏), Huai Opera (淮剧), Jin Opera (晋剧), Peking Opera (京剧), Pingju Opera (评剧), Qin Opera (秦腔), Cantonese Opera (粤剧), Henan Opera (豫剧), Huangmei Opera (黄梅戏), Yue Opera (越剧), Erren Zhuan (二人转), Xi Opera (锡剧), Yang Opera (扬剧), Lu Opera (吕剧), Suzhou Pingtan (苏州评弹), Jingyun Dagu (京韵大鼓).

The above 21 genres selected 31750 tracks for experimentation from tens of thousands of traditional opera pieces. Before the experiment, to ensure audio quality and content reliability, the audio with much noise or without song music [1] was removed. Furthermore, the method of human rating [23] is used to ensure the correctness of the target label. Table 1 gives the statistical data information of each opera class. The total duration includes 7466.85 hours, with an average of 14.11 minutes per track.

Table 1: *Statistical information of the opera class*

| Index | Opera Type | Num | Duration/h |
|---|---|---|---|
| 1 | Teochew Opera | 731 | 397.56 |
| 2 | Kunqu Opera | 503 | 148.68 |
| 3 | Sichuan Opera | 1190 | 346.27 |
| 4 | Hebei Clapper Opera | 1057 | 489.33 |
| 5 | Shanghai Opera | 1583 | 274.02 |
| 6 | Huagu Opera | 540 | 213.0 |
| 7 | Huai Opera | 1305 | 239.47 |
| 8 | Jin Opera | 1123 | 180.6 |
| 9 | Peking Opera | 1350 | 202.22 |
| 10 | Pingju Opera | 1062 | 412.75 |
| 11 | Qin Opera | 1578 | 275.79 |
| 12 | Cantonese Opera | 1587 | 210.27 |
| 13 | Henan Opera | 1159 | 289.53 |
| 14 | Huangmei Opera | 2457 | 840.43 |
| 15 | Yue Opera | 1248 | 471.02 |
| 16 | Erren Zhuan | 2797 | 556.38 |
| 17 | Xi Opera | 881 | 193.81 |
| 18 | Yang Opera | 1103 | 181.99 |
| 19 | Lu Opera | 2948 | 551.25 |
| 20 | Suzhou Pingtan | 5288 | 951.61 |
| 21 | Jingyun Dagu | 260 | 40.87 |
| Total | 21 Types | 31750 | 7466.85 |

# 5. Experiments and results

## 5.1. Dataset and training detail

Our experimental data were collected in two time periods. The first collected data is used to divide the training and validation sets, while the second collected data is used entirely for the test set. To ensure that the length of the track is not too long, the tracks in the first two datasets are evenly divided into 10 minutes of audio clips. Moreover, the audio clip or track with fewer than 20 seconds is discarded for three datasets. As a result, the training, validation, and test sets contain 21037, 5588, and 6301 audio clips, respectively. Because many audio clips have different channels and sampling rates, we convert all audio clips to monophonic and resample them to 16 kHz.

In this work, the Transformer we used has an embedding dimension of 256, one layer, and eight heads, where positional embedding is unnecessary, and the CLS token is optional. The Mel-spectrogram of a randomly sampled 15 s audio clip is computed with 128 mel filter banks, 512 samples of Hann window, and a hop size of 256 samples which is utilized as input to the proposed network. Furthermore, we minimize binary cross-entropy loss and update trainable parameters using a mixture of scheduled ADAM [24] and stochastic gradient descent (SGD) during training. In the inference stage, average track predictions are performed to get the final prediction.

Because opera has similar musical expression to music, we investigate several methods in [14] for our task. They are FCN [25], Musicnn, Sample-level [26], Sample-level with squeeze-and- excitation (SE) [27], CRNN [28], CNNSA [15], Harmonic CNN [29], Short-chunk CNN [14], Short-chunk CNN with residual connections (RES) [14] and AST. Among them, CNNSA is selected as our baseline model. In the rest of the paper, we will denote the proposed joint time and frequency transformer with criss-cross attention as JTFT-CCA and with factorised self-attention as JTFT-FSA.

Except for Accuracy, F1, Precision, and Recall, Area Under Precision Recall Curve (PR-AUC) and Area Under Receiver Operating Characteristic curve (ROC-AUC) are also used as Evaluation Metrics.

## 5.2. Results

To compare the performance of different models, we report evaluation metrics of all implemented models using the test dataset in Table 2. The results in the table above demonstrate that our proposed JTFT-CCA achieves the best classification results in all metrics, and the model with the waveform as the input gets slightly worse results. From both columns of parameters and evaluation metrics, we find that our model has a smaller number of parameters in achieving optimal classifica-

Table 2: *Results of different models on the test dataset*

| Models | #param | Acc | F1 | P | R | ROC-AUC | PR-AUC |
|---|---|---|---|---|---|---|---|
| FCN[25] | 0.45m | 87.58% | 83.72% | 85.64% | 85.09% | 0.9909 | 0.9189 |
| Musicnn[13] | 0.78m | 75.24% | 69.98% | 75.48% | 72.56% | 0.9741 | 0.8077 |
| Sample-level[26] | 1.86m | 76.96% | 73.74% | 77.81% | 75.80% | 0.9766 | 0.8278 |
| Sample-level + SE[27] | 6.94m | 82.25% | 80.26% | 83.33% | 81.48% | 0.9838 | 0.8890 |
| CRNN[28] | 0.39m | 88.14% | 85.65% | 86.63% | 86.76% | 0.9920 | 0.9296 |
| Harmonic CNN[29] | 3.62m | 88.55% | 86.14% | 87.44% | 86.94% | 0.9908 | 0.9370 |
| Short-chunk CNN[14] | 3.67m | 84.84% | 80.58% | 82.35% | 82.04% | 0.9884 | 0.9045 |
| Short-chunk CNN + RES[14] | 12.09m | 84.11% | 79.36% | 81.11% | 81.03% | 0.9852 | 0.8865 |
| AST[7] | 87.74m | 76.31% | 73.38% | 79.35% | 74.31% | 0.9714 | 0.8471 |
| CNNSA[15] | 10.51m | 91.21% | 88.75% | 88.76% | 89.83% | 0.9946 | 0.9520 |
| JTFT-FSA (ours) | 2.90m | 90.16% | 86.99% | 87.31% | 88.23% | 0.9927 | 0.9420 |
| JTFT-CCA (ours) | 2.57m | **92.51%** | **90.57%** | **90.91%** | **91.45%** | **0.9958** | **0.9695** |

Table 3: *Ablation Results*

| Models | #param | Acc | F1 | P | R | ROC-AUC | PR-AUC |
|---|---|---|---|---|---|---|---|
| JTFT-CCA | 2.57m | **92.51%** | **90.57%** | **90.91%** | **91.45%** | **0.9958** | **0.9695** |
| Time-only | 2.39m | 89.78% | 86.77% | 87.37% | 87.82% | 0.9936 | 0.9423 |
| Frequency-only | 2.39m | 88.95% | 85.59% | 86.11% | 86.91% | 0.9925 | 0.9343 |

tion results. Comparing the results of JTFT-CCA and CNNSA, it can be concluded that that frequency information improves the classification results. Furthermore, AST that performed well on music classification instead achieved poorer results on opera classification, indicating that opera has higher complexity than music.

To observe the model's classification performance on genre, Figure 2 gives the confusion matrix for all classes. Each row shows whether the tracks from a genre are misclassified to other genres, and each column indicates the misclassification of other genres on that genre. Those on the diagonal are correctly classified.

From the diagonal values in Figure 2, most genres can be correctly classified. Looking at each row of data, the model easily misclassifies Huagu Opera, Henan Opera and Yang Opera into other genres. Furthermore, observing the column data, we find that the model tends to misclassify the other genres into Huagu Opera, Jin Opera, Huangmei Opera, Erren Zhuan, Yang Opera, and Lu opera. Among them, Huagu Opera and Yang Opera have poor data performance on both rows and columns.

### 5.3. Ablation study

To verify the rationality of the proposed model, we extend our previous comparisons by doing an ablation study on our method. We evaluate whether using time and frequency attention on transformer blocks improves the performance. We do so by alternately removing time or frequency attention in Transformer. Specifically, Time-only calculates only time attention while combining frequency dimension with batch size. Instead, Frequency-only applies a similar approach as above to the frequency dimension. Finally, the result is suggested in Table 3. From the results, both Frequency-only and Time-only are worse than JTFT-CCA. It illustrates that time and frequency information is helpful for classification under this task.

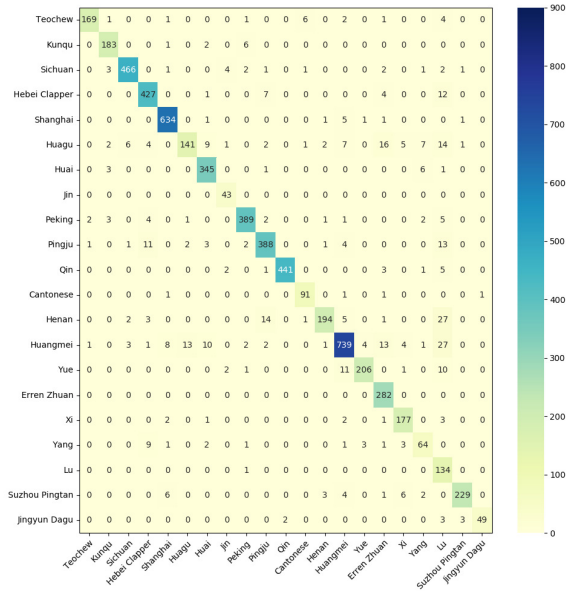


Figure 2: *Confusion matrix of the proposed JTFT-CCA model under the test dataset.*

## 6. Conclusion and future work

In this paper, a neural network that combines both convolution and Transformer is proposed for Chinese opera classification. Notably, a different way to compute the time and frequency attention in a Transformer is used and compared in this paper. The experiments demonstrate that the proposed model achieves optimal performance with fewer model parameters. In future work, this network will be used to distinguish different singing styles in the same genre.

# 7. References


[1] R. Islam, M. Xu, and Y. Fan, "Chinese traditional opera database for music genre recognition," in *2015 International Conference Oriental COCOSDA held jointly with 2015 Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE)*, Shanghai, China, Dec. 2015, pp. 38–41.

[2] R. C. Repetto, R. Gong, N. Kroher, and X. Serra, "Comparison of the singing style of two Jingju schools," in *Pro. ISMIR 2015 – 16th International Society for Music Information Retrieval Conference*, Malaga, Spain, Oct. 2015, pp. 507–513.

[3] Y. Yang, "Structure analysis of beijing opera arias," Ph.D. dissertation, Master Thesis, Universitat Pompeu Fabra, Barcelona (Spain), 2016.

[4] A. Srinivasamurthy, R. Caro Repetto, H. Sundar, and X. Serra, "Transcription and recognition of syllable based percussion patterns: The case of Beijing Opera," in *Pro.ISMIR 2014 – 15th International Society for Music Information Retrieval Conference*, Taiwan, China, Oct. 2014, pp. 431–436.

[5] S. Zhang, R. Caro Repetto, and X. Serra, "Study of the similarity between linguistic tones and melodic pitch contours in Beijing opera singing," in *Pro.ISMIR 2014 – 15th International Society for Music Information Retrieval Conference*, Taiwan, China, Oct. 2014, pp. 343–348.

[6] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *Pro. ICLR 2021 – 9th International Conference on Learning Representations*, Vienna, Austria, May 2021. [Online]. Available: https://openreview.net/pdf?id=YicbFdNTTy

[7] Y. Gong, Y.-A. Chung, and J. Glass, "AST: Audio Spectrogram Transformer," in *Pro. INTERSPEECH 2021 – 22nd Annual Conference of the International Speech Communication Association*, Brno, Czechia, Aug. 2021, pp. 571–575.

[8] G. Bertasius, H. Wang, and L. Torresani, "Is space-time attention all you need for video understanding?" in *Pro. ICML 2021 – 38th International Conference on Machine Learning*, Vienna, Austria, Jul. 2021, pp. 813–824.

[9] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić, and C. Schmid, "Vivit: A video vision transformer," in *Pro. ICCV 2021 – International Conference on Computer Vision*, Montreal, Canada, Oct. 2021, pp. 6836–6846.

[10] Y.-B. Zhang, J. Zhou, and X. Wang, "A study on Chinese traditional opera," in *Proc. ICMLC 2008 – International Conference on Machine Learning and Cybernetics*, Kunming, China, Jul. 2008, pp. 2476–2480.

[11] J. Wang, C. Wang, J. Wei, and J. Dang, "Chinese opera genre classification based on multi-feature fusion and extreme learning machine," in *Proc. APSIPA 2015 – Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, Hong Kong, China, Dec. 2015, pp. 811–814.

[12] H. Zhang, Y. Jiang, W. Zhao, T. Jiang, P. Hu, and T. M. Entertainment, "Chinese opera genre investigation by convolutional neural network," in *Pro.ISMIR 2021 – 22nd International Society for Music Information Retrieval Conference*, Nov. 2021. [Online]. Available: https://ismir2021.ismir.net/lbd/

[13] J. Pons, O. Nieto, M. Prockup, E. Schmidt, A. Ehmann, and X. Serra, "End-to-end learning for music audio tagging at scale," in *Pro.ISMIR 2018 – 19th International Society for Music Information Retrieval Conference*, Paris, France, Sep. 2018, pp. 637–644.

[14] M. Won, A. Ferraro, D. Bogdanov, and X. Serra, "Evaluation of cnn-based automatic music tagging models," in *Pro.SMC 2020 – 17th Sound and Music Computing*, Torino, Italy, Jun. 2020, pp. 331–337.

[15] M. Won, S. Chun, and X. Serra, "Toward interpretable music tagging with self-attention," *arXiv preprint arXiv:1906.04972*, 2019.

[16] M. Won, K. Choi, and X. Serra, "Semi-supervised music tagging transformer," in *Pro.ISMIR 2021 – 22nd International Society for Music Information Retrieval Conference*, Nov. 2021, pp. 769–776. [Online]. Available: https://doi.org/10.5281/zenodo.5624405

[17] W.-T. Lu, J.-C. Wang, M. Won, K. Choi, and X. Song, "Spectnt: A time-frequency transformer for music audio," in *Pro.ISMIR 2021 – 22nd International Society for Music Information Retrieval Conference*, Nov. 2021, pp. 769–776. [Online]. Available: https://doi.org/10.5281/zenodo.5624503

[18] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu, "Ccnet: Criss-cross attention for semantic segmentation," in *Pro. ICCV 2019 – International Conference on Computer Vision*, Seoul, Korea, Oct. 2019, pp. 603–612.

[19] A. Bulat, J. M. Perez Rua, S. Sudhakaran, B. Martinez, and G. Tzimiropoulos, "Space-time mixing attention for video transformer," in *Proc. NeurIPS 2021 – 35th Conference on Neural Information Processing Systems*, Dec. 2021, pp. 19 594–19 607.

[20] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *Proc. ECCV 2016 – 14th European Conference on Computer Vision*, Amsterdam, The Netherlands, Oct. 2016, pp. 630–645.

[21] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. NIPS 2017 – 31st Annual Conference on Neural Information Processing Systems*, Long Beach, CA, USA, Dec. 2017, pp. 5998–6008.

[22] Q. Zhou and L. Lu, *Intermediate Tourism Culture Practice*. Beijing, China: Tsinghua University Press, 2011.

[23] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *Proc. ICASSP 2017 – IEEE International Conference on Acoustics, Speech and Signal Processing*, New Orleans, LA, USA, Mar. 2017, pp. 776–780.

[24] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Pro. ICLR 2015 – 3rd International Conference on Learning Representations*, San Diego, CA, USA, May 2015.

[25] K. Choi, G. Fazekas, and M. Sandler, "Automatic tagging using deep convolutional neural networks," in *Pro.ISMIR 2016 – 17th International Society for Music Information Retrieval Conference*, New York, USA, Aug. 2016, pp. 805–811.

[26] J. Lee, J. Park, L. Kim, and J. Nam, "Sample-level deep convolutional neural networks for music auto-tagging using raw waveforms," in *Pro.SMC 2017 – 14th Sound and Music Computing*, Espoo, Finland, Jul. 2017, pp. 220–226.

[27] T. Kim, J. Lee, and J. Nam, "Sample-level cnn architectures for music auto-tagging using raw waveforms," in *Proc. ICASSP 2018 – IEEE International Conference on Acoustics, Speech and Signal Processing*, Calgary, AB, Canada, Apr. 2018, pp. 366–370.

[28] K. Choi, G. Fazekas, M. Sandler, and K. Cho, "Convolutional recurrent neural networks for music classification," in *Proc. ICASSP 2017 – IEEE International Conference on Acoustics, Speech and Signal Processing*, New Orleans, LA, USA, Mar. 2017, pp. 2392–2396.

[29] M. Won, S. Chun, O. Nieto, and X. Serrc, "Data-driven harmonic filters for audio representation learning," in *Proc. ICASSP 2020 – IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, May 2020, pp. 536–540.