Research Article

Multimodal Music Emotion Recognition Method Based on the Combination of Knowledge Distillation and Transfer Learning

Guiying Tong

College of Music, Huizhou University, Huizhou 516000, Guangdong, China

Correspondence should be addressed to Guiying Tong; tgy@hzu.edu.cn

Received 11 December 2021; Revised 4 January 2022; Accepted 7 January 2022; Published 11 February 2022

Academic Editor: Baiyuan Ding

Copyright © 2022 Guiying Tong. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

The main difficulty of music emotion recognition is the lack of sufficient labeled data. Only the labeled data with unbalanced categories are used to train the emotion recognition model. Not only is accurate labeling of emotion categories costly and time-consuming, but it also requires extensive musical background for labelers. At the same time, the emotion of music is often affected by many factors. Singing methods, music styles, arrangement methods, lyrics, and other factors will affect the expression of music emotions. This paper proposes a multimodal method based on the combination of knowledge distillation and music style transfer learning and verifies the effectiveness of the method on 20,000 songs. Experiments show that compared with traditional methods, such as single audio, single lyric, and single audio with multimodal lyric methods, the method proposed in this paper has significantly improved the accuracy of emotion recognition, and the generalization ability has been significantly improved.

1. Introduction

Music works contain rich human emotions, and emotions play an indispensable role in the transmission of musical emotions and understanding and appreciation of music [1–3]. With the current development of Internet technology and artificial intelligence, the amount of digital music is growing rapidly. With a large number of music works, how to recommend suitable music according to different environments and different moods of users has become a hot topic of research in recent years. In this context, the automatic recognition technology of music emotion has attracted more and more attention from the industry. In recent years, deep learning technology has replaced traditional statistical algorithms as the mainstream technology in the field of automatic music emotion recognition [4–6]. The main content of music includes digital audio and lyrics text, and the current research on music emotion recognition mainly focuses on these two aspects.

Existing studies have applied artificial intelligence and Internet of Things technology to the teaching of art courses and the work of art exhibition. In traditional statistical methods, models need to design complex elements and

features for repeated understanding and recognition manually, which requires a lot of time and learning costs. Unlike traditional methods, deep learning algorithms can automatically identify and extract the most appropriate representative feature elements from the data. At the same time, deep learning has shown outstanding contributions in many fields such as speech recognition and machine translation. Some scholars have achieved excellent results in the music emotion classification competition task through the in-depth application of deep learning technology (Music Information Retrieval Evaluation Exchange), combined with the emotion recognition method of music audio [7–9] in Figure 1.

Some scholars also proposed an end-to-end multimodal emotion recognition binary classification method based on music audio and lyrics, and the superiority of this method was verified through specific experiments. Compared with the emotion recognition method of single music audio and single lyric music, this method has a significant improvement in recognition accuracy [10–12]. Some scholars have conducted research on the emotion recognition of multimodal music and carried out detailed comparative experiments on the methods of fusion between different music in

FIGURE 1: The structure of the student-teacher model.

different periods. The results show that the middle fusion method of music is better than the front fusion and the post-fusion. In addition, some researches have carried out comparative analysis and research on several classic music emotion data sets that are currently available in the industry, e.g., CAL500, MIREX 2007 AMC, and MediaEval Emotion in Music. It is found that labeling music emotion data sets has high cost and is relatively time-consuming. In the research on the related content of the music public data set, it is found that the data quantity and quality of the public data set in this respect are not ideal, and many data sets have problems such as the unbalance of emotion categories [13–15].

Based on the multimodal music emotion recognition method, combining knowledge distillation and transfer learning, this article attempts to improve the accuracy of music emotion recognition effectively when the amount of labeled data is too small or the emotion categories are not balanced. The method based on knowledge distillation specifically analyzes the application scenarios of the teacher-student model in the image field. On the one hand, the model shows relatively excellent processing performance [16–18]. On the other hand, it also builds a multimodal neural network structure based on audio and lyrics recognition methods. This network structure can quickly identify and efficiently classify different music styles, which is conducive to the rapid application of music recommendation [19–23] models in Figure 2.

2. Data Preprocessing

2.1. Audio Signal Expression and Preprocessing Technology. Mel spectrum, as a signal representation method generally applicable to common audio classification tasks, is better than other high-level audio signal representations. This method relatively completely retains the characteristic information of a series of signals in the music work and truly achieves the lossless preservation of the music information [24–27]. At the same time, Mel's frequency spectrum is more in line with the auditory characteristics of the field of ergonomics. Based on the above multiple factors, this paper selects the Mel spectrum method as the input data for music

audio analysis. In addition, this article adopts the Voice Activity Detection (VAD) method to specifically detect whether there is a silent frame in the music signal. Considering the audio part of the silent frame will affect the overall recognition of the specific audio and the flow of the expression and preprocessing of the music audio signal in this paper shown in Figure 3.

A piece of music mainly includes audio information and lyrics text information, and audio information includes time domain features, frequency domain features, and cestrum features. Time domain features usually refer to time domain parameters calculated in fixed-length music, which mainly include short-term energy, short-term average amplitude, and average amplitude difference functions. Frequency domain characteristics refer to the characteristic parameters obtained by converting fixed-length music from time domain signals into frequency domain signals through Fourier transform, which mainly include spectrum centroid, spectrum roll-off, and spectrum flux. The cestrum feature refers to the use of human ears to have different perception effects on the different loudness, pitch, and timbre of the sound. Therefore, it is used to classify different music emotions. The specific operation methods are as follows: (1) the model divides the music into specific lengths and uses Fourier conversion to frequency domain signals; (2) the model calculates the logarithm of the frequency domain signals and then performs the inverse Fourier transform [28–31].

2.2. Word Vector Pre-Representation of Lyrics Text.

Taking into account the types of English songs used in this article, the lyrics of the songs include English and other different types of languages. In order to ensure the compatibility of the system with multiple languages, this article cuts the words of each character in English songs according to special characters, which guarantees the versatility and fluency of the word-slicing algorithm [32–35]. The word vector expresses the adaptive expression method of dynamic lyrics recognition and segmentation. At the same time, this paper performs word frequency statistics and word vector initialization processing on the pretrained lyrics collection. In the specific

FIGURE 2: Multimodal neural network structure based on audio and lyrics text.

FIGURE 3: System pretreatment process.

model training, for the convenience of adjusting the word vector, part of the algorithm optimization work is carried out [36–38]. In addition, this article sets the word vector word_dim to 138, and the maximum length parameter of lyrics to 200. In this paper, the preprocessing of the word vector in the lyric text is shown in Figure 4.

3. Student-Teacher Model and Transfer Learning Methods

3.1. Student-Teacher Model. Considering that the data was less labeled and relatively unevenly distributed, this paper uses the process model of Figure 1 to express the student-teacher model, thereby improving the accuracy of the overall music emotion recognition. The teacher-student model specifically uses the existed music genre network recognition architecture, and we then use different genre recognition systems to learn the network parameters in stages.

At the same time, considering that the network reasoning ability and performance of the teacher system are generally higher than those of the student group, this paper specifically optimizes the network parameters of the music style selected in the teacher model, and its performance is generally better than the network parameters of the students. On this basis, the characteristics, advantages, and disadvantages of this type of network structure were further elaborated.

The machine learning method used in the text mainly uses the thinking mode of transfer learning and mainly uses two methods for knowledge transfer. The first mode is to use the trained song style network structure as a feature extraction tool for emotion recognition and use it for the recognition of part of the music emotion network after feature training. The second mode mainly combines the already trained song style network structure with the newly added network into the system for training. This article mainly chooses the second training mode. Considering that the teacher network model does not participate in the backpropagation of the neural network system, this paper compares the teacher model parameter W with the student model parameter w through Exponential Moving Average (EMA) analysis; the specific parameter expression results of the teacher model at time t are as follows:

The Wt expression of the most commonly used function is

$$\begin{aligned} Wt &= \alpha \times Wt - 1 + (1 - \alpha) \times wt, \\ \text{Floss1} &= Hw(\alpha) + \lambda Kw * Lw(\alpha), \\ \text{Floss2} &= Hw(\alpha) + \lambda w * KL(\alpha) + (1 - \alpha) \times wt, \\ \text{Floss3} &= Hw(\alpha) + \lambda w * KL(\alpha) + (1 - \alpha) \times wt + \alpha \\ &\quad \times Wt - 1. \end{aligned} \quad (1)$$

Among them, α represents the attenuation-smoothing rate, Wt represents the parameters of the student model at

FIGURE 4: Word vector pre-expression of lyrics.

time t . The F-loss function of the teacher-student model is composed of relative entropy (Kullback–Leibler divergence) $KL(\alpha)$ and classification cross entropy $Hw(\alpha)$ (control of the importance of two parts of CE). For the same sample X , first get the output s through the student model, then get the output t through the teacher model, and use the data t and s to calculate the $KL(\alpha)$ loss. On this basis, the result data of the gradient analysis calculation is passed back to the student model, and the CE loss value in the s model is further calculated. Finally, according to the two parts of the loss function, the specific parameters of the student model are updated; the student model parameters are further analyzed by exponential smoothing.

At the same time, in the process of model training, iterative analysis of model parameters is required. Input the sample data set X into the model. The data set contains both emotionally labeled samples and unlabeled samples. In addition, these sample sets are enhanced through the system model, which further improves the generalized understanding of emotional data of the system machine learning model.

3.2. Data Enhancement

3.2.1. Gaussian Noise Analysis. English songs are often accompanied by various types of noise during the specific recording and dissemination process. In the process of preprocessing the audio, this paper can further use Gaussian noise processing for all audio content and perform noise enhancement processing on audio effects.

3.2.2. Audio Cutting Process. Before analyzing the audio, perform audio tailoring according to different audio types and styles. According to different types of song fragments, the machine learning model system simulates human emotions when listening to music. The average length of the audio clip collection of the model is set to 30 seconds to ensure the stability and fluency of the audio recognition process of the system.

3.2.3. Audio Mixing Analysis. This article randomly mixes different types of audio. The model specifically assumes that a pair of audio samples (X_a , X_b) are processed, and the resulting mixed sample is set to X_{com} . It should be noted

that the generated X_{com} is only used to calculate the loss of the model KL and cannot be used for multidimensional systematic mixing of different types of audio for emotional tags. Among them, γ is the sample-mixing coefficient.

$$\begin{aligned} X_{com} &= (1 - \gamma) * X_a + \gamma * X_b, \\ Y_{com} &= (1 - \gamma) * Y_a + \gamma * Y_b. \end{aligned} \quad (2)$$

Forward propagation of working signal: take a sample N_m from the sample set and set θ_i as the threshold of the i neuron of the hidden layer:

$$\begin{aligned} Y_m &= f \left(\sum_{i=1}^m w_{ih} X_m - \theta_i \right), \\ O_j &= g \left(\sum_{j=1}^m w_{hj} Y_m - \theta_j \right). \end{aligned} \quad (3)$$

In addition to audio information in music, lyrics text information also implies certain emotions. In order to avoid discretization of emotional expression due to the sparse distribution of words in the lyrics, short sentences, Y_m , and repetition, it is necessary to filter the lyrics. First, create a vocabulary of word emotion recognition. The common words in the lyrics are classified according to the emotions and strengths they express. For example, the term “sun” is usually used to express hopeful and positive emotions, while the term “car” expresses emotions that are rather vague. Therefore, in terms of emotion recognition, “sun” is more efficient than “car.” When creating a vocabulary of word emotion recognition, vocabulary like “sun” should be kept, and vocabulary like “car” should be avoided at the same time. After the word emotional recognition vocabulary is created, O_j , a certain vocabulary, needs to be identified and quantified. The specific formula is as follows. The total network error is

$$e_{(n)} = \frac{1}{2} \sum_{j=1}^m (O_j - d_j)^2. \quad (4)$$

Because the data collected in the research are not in the same order of magnitude, the prediction error of neural network will become larger, and even the network output value will fall into a certain interval, it is impossible for $e(n)$ to carry out normal output and error feedback. Therefore,

this paper normalizes the original data $f(x)$ to eliminate the data dimension. In this paper, map-minimax ($\max_x - \min_x$) is used to process the original data into data in $[-1, 1]$ interval:

$$\begin{aligned} y &= (\max y - \min y) \frac{x - \min x}{\max x - \min x} + \min y, \\ x &= (\max x - \min x) \frac{y - \min y}{\max y - \min y} + \min x, \\ f(x) &= (1 + e^{-ax})^{-1}, \\ f(x) &= \frac{1 - e^{-x}}{1 + e^{-x}}, (-1 < f(x) < 1). \end{aligned} \quad (5)$$

4. Multimodal Network Structure

The teacher-student model proposed in this article specifically uses a multimodal neural network structure based on text information such as audio and lyrics. Specifically, the input audio will be cut into N segments according to a fixed length. The model outputs the final sentiment label as N segments and averagely obtains its analysis results.

4.1. Audio Model. This article specifically uses the Mel spectrum information used in the previous article as the model data parameters. The Mel spectrum model is a two-dimensional matrix data with time dimension and characteristic length. Specifically, the time length `feature_len` = 1,024, and the feature length `mel_dim` = 128, so the number of input samples for the overall model is $1,024 \times 128$. In view of the relatively little emotional information in the lyrics, this paper compares and analyzes the model with the existing network analysis models such as ResNet, GoogLeNet, and VGGNet. Specifically, this paper designs a relatively suitable lightweight convolutional neural network system and then sorts out the network organization structure of the audio model, as shown in Figure 5. The convolutional neural network selected in this article has a certain feasibility, considering that many objects in life have local correlations. In particular, each frame of signal in a music piece is not isolated but is linked and interacts with others, so as to effectively convey the music information to the listener.

In order to prevent words with strong emotional inclinations and less frequent occurrences, they are filtered due to the low results of the above formula. Therefore, improvements are needed for such words. This article uses synonyms to prevent such words from being eliminated. The specific idea is to find the similar words to the words in the vocabulary and add them to the vocabulary after creating a vocabulary of word emotion recognition. The process flow of multifeature fusion of audio information and lyric text is shown in the Figure 5. Preprocessing is required before feature extraction. The preprocessing of audio information mainly includes signal filtering. The preprocessing of lyrics text is the operation of word emotion recognition vocabulary

and supplementing similar words. After preprocessing, feature extraction is performed separately; that is, feature fusion and output are performed. Aiming at the feature vector data of music audio and lyrics, this paper improves the traditional forward neural network. Using specific Chebyshev orthogonal polynomial clusters, a forward neural network model with a specific structure is constructed. In this model, the forward neural network uses a single hidden layer to reduce the complexity of the overall model.

The activation function of each neuron in the hidden layer uses each function in Chebyshev's orthogonal polynomial cluster, while the activation functions of the neurons in other layers of the model all use linear activation functions. The weight of each neuron in the hidden layer and the output layer is w . The specific structure of the model is shown in the figure below. Emotion recognition classification can effectively improve the quality of music retrieval services and has important practical significance. In this paper, the Chebyshev polynomial cluster is incorporated into the forward neural network model, and the recognition of higher music emotions is achieved under the supervision and training of the gradient descent learning algorithm. It is noted that the accuracy of music sentiment classification is not only related to the number of layers of the model and the number of neurons, but also related to the number of dimensions of training samples.

4.2. Text Model. This paper focuses on screening the song's lyrics information, song title information, album style, and other characteristic information. Specifically, the maximum text length `max_length` is set to 200, the word vector dimension parameter `word_dim` is set to 128, and the word vector matrix-embedding matrix of the lyrics text is set to 200×128 matrix data. Specifically, the initial value of embedding matrix refers to the above, and the matrix data of the word vector is further incorporated into the model to train the stability and recognition accuracy of the model. The network architecture in this model is shown in Figure 6.

4.3. Multimodal Model. The text uses a multimodal model to combine the text model and the individual audio and then remove the respective full link parts. On this basis, the text model and the audio model are combined and merged, and then the emotional label prediction results are divided into N equal parts. The specific structure mode of the model is shown in Figure 6.

5. Experimental Results and Analysis

Considering that the previous article has analyzed the existing methods of multimodal emotion recognition, compared with traditional single audio or lyrics methods, the multimodal emotion model used in this article has significant advantages. In addition, the experiments in this article do not pay attention to the actual comparison of these two types of methods. The model in this paper mainly verifies the effect of knowledge distillation and transfer learning algorithms on existing networks under the

FIGURE 5: Audio model network.

FIGURE 6: Multimodal network structure.

condition of missing or unbalanced sentiment labeled data. The data set used in the experiment contains about 2,000 English sentiments with or without annotated emotions.

5.1. Introduction to Related Data Sets. The data set used in this experiment contains about 2,000 English songs. In the repertoire, there are about 87% of English songs and about 13% of songs in other words. In addition, the data set comprehensively selects 15,000 songs as the specific training set and 5,000 songs as the test data. The emotion of the song is subdivided into four types: sad, lyrical, happy, and not sad or not happy. The music emotion used in this article expresses a relatively discrete model, and each category is mainly labeled with mutually exclusive discrete states. The number of songs in each category in the training set is 6,000 (2,000 labeled; 4,000 unlabeled), 7,000 (4,000 labeled; 3,000 unlabeled), and 2,000 (800 labeled; 1,200 unlabeled). Among them, the audio uses a unified WAV format, the sampling frequency is 2,205 Hz, and the monophonic channel is sampled. The music data was labeled by 5 people with more than 5 years of music education experience, and the samples of more than 3 people were included in the sample database. The data labeling of the training set is shown in Figure 7.

5.2. Comparison of Model Methods. Experiment 1, experiment 2, and experiment 3 compare and analyze the fitting results of different models. Specifically, this paper compares the results produced by two experimental methods based on knowledge distillation and nonknowledge distillation. Considering that in different experimental environments, the number of labeled data items in the knowledge distillation method is small and unbalanced, the recognition accuracy of its training set and test set has shown a significant improvement in different models. The experimental results are shown in the figure below. The results indicate that using the teacher-student model combined with the knowledge distillation method for analysis, the teacher network can guide the student network to learn the answers consistent with the teacher network from the marked and unmarked emotional data. At the same time, the teacher network model is relearned according to the model parameters completed in different stages of the student network training, and the best model parameters are obtained after continuous iteration. The research results show that the accuracy of the multimodal method combined with the knowledge distillation method is greatly improved compared with the uncombined multimodal method in Figures 8 and 9.

In experiment 2, a comparative analysis of the knowledge distillation multimodal method combined with genre transfer learning is conducted.

Furthermore, the multimodal method combining only knowledge distillation and the model method proposed are compared and analyzed in this paper. Overall, this paper adopts the knowledge distillation multimodal method for song transfer learning and compares the model learning results with traditional multimodal learning methods. The experimental results are shown in Figure 10. The results show

that when the initial values of the parameters of the model are selected, we further substitute the data into the model to analyze whether the local optimal solution can be obtained. In the traditional case, the model may not be able to analyze the global optimal solution. Existing experimental analysis further shows that the use of this paper's genre transfer algorithm can identify the key characteristics of music emotions and perform transfer learning. The model has high generalization adaptability. The initial values of the model parameters generally affect whether the model falls into a local optimal solution and even cause the model to be unable to solve the global optimal solution.

Through experimental analysis, it can be found that the genre characteristics of songs can be well identified in the model. Further transfer learning in the process of music emotion model recognition can show the model's good generalization ability. At the same time, the model achieves the global optimal model effect very well and converges to a higher accuracy rate in the case of fewer iterations in Figures 11 and 12.

Experiment 3 includes a comparison between the lightweight convolutional neural network and the classic network in the audio model selected in this article.

This paper avoids the use of words with strong emotional tendencies, so as to avoid biased analysis results. Improvements are needed for such words. This article uses synonyms to prevent such words from being eliminated. The specific idea is to find the similar words to the words in the vocabulary and add them to the vocabulary after creating a vocabulary of word emotion recognition. The process flow of multifeature fusion of audio information and lyric text is shown in the figure below. Preprocessing is required before feature extraction. The preprocessing of audio information mainly includes signal filtering. The preprocessing of lyrics text is the operation of word emotion recognition vocabulary and supplementing similar words. After preprocessing, feature extraction is performed separately; that is, feature fusion and output are performed. Aiming at the feature vector data of music audio and lyrics, this paper improves the traditional forward neural network. Using specific Chebyshev orthogonal polynomial clusters, a forward neural network model with a specific structure is constructed. In this model, the forward neural network uses a single hidden layer to reduce the complexity of the overall model.

This paper compares the selected lightweight convolutional neural network with the classic network, mainly to illustrate that under the actual data set, choosing a suitable design of the lightweight convolutional neural network is better than the generalization ability brought by the classic network promotion. This article compares the two without combining knowledge distillation and transfer learning to ensure the independence of experimental comparison. The experimental results are shown in Figure 13. The method of using the classic ResNet network on the training set shows a higher accuracy rate, but on the test set its accuracy rate drops more. According to this conclusion, the model shows that there is an overfitting phenomenon under the ResNet network model, and it cannot have the generalization ability on the entire data set in Figures 14 and 15.

FIGURE 7: Statistical table of data labeling of control group set.

FIGURE 8: Result statistics of experiment 1.

FIGURE 9: Result statistics of multimodal music emotion recognition in experiment 2 in a week.

FIGURE 10: Result statistics of multimodal music emotion recognition in experiment 3.

FIGURE 11: Result statistics of multimodal music emotion recognition in experiment 4 in a week.

FIGURE 12: Comparison of conclusions of different data models.

FIGURE 13: Experimental comparison results using different convolutional networks.

FIGURE 14: Result statistics of multimodal music emotion recognition in experiment 4.

FIGURE 15: The results based on the combination of knowledge distillation.

6. Conclusion

This paper proposes a multimodal method based on the combination of knowledge distillation and music style transfer learning. This research uses the knowledge distillation teacher-student model to unearth the consistent relationship between the unlabeled and labeled emotional data

between the teacher model and the student model. In addition, this study improves the accuracy of the model when there are less sentiment labeled data and unbalanced categories. At the same time, this article uses big data and machine learning analysis methods to make the model learn the characteristics of song genre, which characterizes the transfer learning process in music emotion recognition. In

the case of a small number of iterations, the overall accuracy is higher, and the convergence speed in the training process is improved. The research results prove that the relevant features in the audio field have good effects in emotion recognition tasks. Combining the needs of the research conclusions of this article, two issues need to be explored in the next step of the research. The first is to analyze the ambiguity of sentiment labels in different subjects and analyze the instability of training and testing caused by the ambiguity. The second is to study the influence of big data migration learning for different tasks on the effect of emotion recognition, such as the influence of visual tasks on the effect of emotion recognition.

Data Availability

The data set can be accessed upon request.

Conflicts of Interest

The author declares that there are no conflicts of interest.

Acknowledgments

This work was supported by the research project on 2021 teaching quality and reform of Huizhou University, name: Research on Guzhen Teaching Mode Reform Based on Application-Oriented Talent Cultivation, no. X-JYJG2021070.

References

- [1] M. Doyran, A. Schimmel, P. Baki et al., "Mumbai: multi-person, multimodal board game affect and interaction analysis dataset," *Journal on Multimodal User Interfaces*, vol. 15, no. 4, pp. 373–391, 2021.
- [2] M. C. Camacho, E. M. Williams, K. Ding, and S. B. Perlman, "Multimodal examination of emotion processing systems associated with negative affectivity across early childhood," *Developmental Cognitive Neuroscience*, vol. 48, no. 1, pp. 915–917, 2021.
- [3] D. Liu, L. Chen, Z. Wang, and G. Diao, "Speech expression multimodal emotion recognition based on deep belief network," *Journal of Grid Computing*, vol. 19, no. 2, pp. 59–72, 2021.
- [4] T. Chen, H. Yin, X. Yuan, Y. Gu, F. Ren, and X. Sun, "Emotion recognition based on fusion of long short-term memory networks and SVMs," *Digital Signal Processing*, vol. 117, no. 1, pp. 137–153, 2021.
- [5] D. S. Cortes, C. Tornberg, T. Bnzigier, H. A. Elfenbein, H. Fischer, and P. Laukka, "Effects of aging on emotion recognition from dynamic multimodal expressions and vocalizations," *Scientific Reports*, vol. 11, no. 1, 2021.
- [6] M. Dahmane, J. Alam, P.-L. St-Charles, M. Lalonde, K. Heffner, and S. Foucher, "A multimodal non-intrusive stress monitoring from the pleasure-arousal emotional dimensions," *IEEE Transactions on Affective Computing*, vol. 99, p. 1, 2020.
- [7] L. Schoneveld, A. Othmani, and H. Abdelkawy, "Leveraging recent advances in deep learning for audio-Visual emotion recognition," *Pattern Recognition Letters*, vol. 146, pp. 1–7, 2021.
- [8] S. Hans-Eckhardt, "Music-evoked emotions—current studies," *Frontiers in Neuroscience*, vol. 11, p. 600, 2017.
- [9] M. Tzelepi, N. Passalis, and A. Tefas, "Online subclass knowledge distillation," *Expert Systems with Applications*, vol. 181, pp. 115–132, 2021.
- [10] G. Valle-Pérez, H. Enter G E, J. Beskow, A. Holzapfel, P. Y. Oudeyer, and S. Alexanderson, "Transflower: probabilistic autoregressive dance generation with multimodal attention," *ACM Transactions on Graphics*, vol. 40, no. 6, pp. 1–14, 2021.
- [11] A. Del Ponte, P. Canofari, and A. De Dominicis, "Financial and trade relationships between the Eurozone and China in the age of resilience," *Asia Europe Journal*, vol. 19, no. 4, pp. 489–506, 2021.
- [12] J. Dong, L. Yin, X. Liu, M. Hu, X. Li, and L. Liu, "Impact of internet finance on the performance of commercial banks in China," *International Review of Financial Analysis*, vol. 72, no. 12, pp. 1–12, 2020.
- [13] B. Zhao, K. Kenjegalieva, J. Wood, and A. Glass, "A spatial production analysis of Chinese regional banks: case of urban commercial banks," *International Transactions in Operational Research*, vol. 27, no. 4, pp. 2021–2044, 2019.
- [14] C. Zheng, S. Chen, and Z. Dong, "Economic fluctuation, local government bond risk and risk-taking of city commercial banks," *Sustainability*, vol. 13, no. 17, pp. 1–26, 2021.
- [15] A. G. Assaf, A. N. Berger, R. A. Roman, and M. G. Tsionas, "Does efficiency help banks survive and thrive during financial crises?" *Journal of Banking & Finance*, vol. 106, no. 7, pp. 445–470, 2019.
- [16] O. Badunenko and S. C. Kumbhakar, "Economies of scale, technical change and persistent and time-varying cost efficiency in Indian banking: do ownership, regulation and heterogeneity matter?" *European Journal of Operational Research*, vol. 260, no. 2, pp. 789–803, 2017.
- [17] S. Kumar Dwivedi, R. Amin, S. Vollala, and R. Chaudhry, "Blockchain-based secured event-information sharing protocol in internet of vehicles for smart cities," *Computers & Electrical Engineering*, vol. 86, no. 1, pp. 1–9, 2020.
- [18] Z. Khan and S. Amin, "Bottleneck model with heterogeneous information," *Transportation Research Part B: Methodological*, vol. 112, no. 1, pp. 157–190, 2018.
- [19] J. M. Cairney, K. Rajan, D. Haley et al., "Mining information from atom probe data," *Ultramicroscopy*, vol. 159, no. 1, pp. 324–337, 2020.
- [20] Z. Wang, H. Ren, Q. Shen, W. Sui, and X. Zhang, "Seismic performance evaluation of a steel tubular bridge pier in a five-span continuous girder bridge system," *Structures*, vol. 31, no. 1, pp. 909–920, 2021.
- [21] S. Nakayama and J. Takayama, "Traffic network equilibrium model for uncertain demands," in *Proceedings of the 82nd Transportation Research Board Annual Meeting*, Washington, DC, USA, January 2021.
- [22] H. Shao, W. H. K. Lam, and M. L. Tam, "A reliability-based stochastic traffic assignment model for network with multiple user classes under uncertainty in demand," *Networks and Spatial Economics*, vol. 6, no. 3, pp. 173–204, 2019.
- [23] A. Chen, J. Kim, S. Lee, and Y. Kim, "Stochastic multi-objective models for network design problem," *Expert Systems with Applications*, vol. 37, no. 2, pp. 1608–1619, 2020.
- [24] S.-M. Hosseiniinasab, S.-N. Shetab-Boushehri, S. R. Hejazi, and H. Karimi, "A multi-objective integrated model for selecting, scheduling, and budgeting road construction projects," *European Journal of Operational Research*, vol. 271, no. 1, pp. 262–277, 2018.

- [25] M. Johnson, M. Schuster, Le et al., "Google's multilingual neural machine translation system: enabling zero-shot translation," *Transactions of the Association for Computational Linguistics*, vol. 5, no. 1, pp. 339–351, 2017.
- [26] M. D. Moreno, "Translation quality gained through the implementation of the iso en 17100:2015 and the usage of the blockchain," *Babel*, vol. 1, no. 2, pp. 1–9, 2020.
- [27] X. Wang, X. Yu, L. Guo, F. Liu, and L. Xu, "Student performance prediction with short-term sequential campus behaviors," *Information*, vol. 11, no. 4, p. 101, 2020.
- [28] Q. Guo, Z. Zhu, Q. Lu, D. Zhang, and W. Wu, "A dynamic emotional session generation model based on Seq2Seq and a dictionary-based attention mechanism," *Applied Sciences*, vol. 10, no. 6, pp. 1–10, 2020.
- [29] H. Ren, X. Mao, W. Ma, J. Wang, and L. Wang, "An English-Chinese machine translation and evaluation method for geographical names," *ISPRS International Journal of Geo-Information*, vol. 9, no. 3, pp. 193–201, 2020.
- [30] J. Arús-Pous, T. Blaschke, S. Ulander, J.-L. Reymond, H. Chen, and O. Engkvist, "Exploring the GDB-13 chemical space using deep generative models," *Journal of Cheminformatics*, vol. 11, no. 1, pp. 20–29, 2019.
- [31] T. Moon, T. In Ahn, and E. Son, "Long short-term memory for a model-free estimation of macronutrient ion concentrations of root-zone in closed-loop soilless cultures," *Plant Methods*, vol. 15, no. 1, pp. 1–12, 2019.
- [32] N. Pourdamghani and K. Knight, "Neighbors helping the poor: improving low-resource machine translation using related languages," *Machine Translation*, vol. 33, no. 3, pp. 239–258, 2019.
- [33] L. Bote-Curiel, S. Muñoz-Romero, A. Gerrero-Curieses, and J. Luis Rojo-Álvarez, "Deep learning and big data in healthcare: a double review for critical beginners," *Applied Sciences*, vol. 9, no. 11, pp. 1–11, 2019.
- [34] J. Zhang and T. Matsumoto, "Corpus augmentation for neural machine translation with Chinese-Japanese parallel corpora," *Applied Sciences*, vol. 9, no. 10, pp. 1–12, 2019.
- [35] Y. Chen, Y. Ma, X. Mao, and Q. Li, "Multi-task learning for abstractive and extractive summarization," *Data Science and Engineering*, vol. 4, no. 1, pp. 14–23, 2019.
- [36] X. Xiao, "Analysis on the employment psychological problems and adjustment of retired athletes in the process of career transformation," *Modern Vocational Education*, vol. 5, no. 12, pp. 216–217, 2018.
- [37] Y. Zhou and B. Yang, "Sports video athlete detection using convolutional neural network," *Journal of Natural Science of Xiangtan University*, vol. 39, no. 1, pp. 95–98, 2017.
- [38] J. Pang, "Research on the evaluation model of sports training adaptation based on self-organizing neural network," *Journal of Nanjing Institute of Physical Education*, vol. 16, no. 1, pp. 74–77, 2017.