# Retraction Notice

The Editor-in-Chief and the publisher have retracted this article, which was submitted as part of a guest-edited special section. An investigation uncovered evidence of systematic manipulation of the publication process, including compromised peer review. The Editor and publisher no longer have confidence in the results and conclusions of the article.

JC and HW either did not respond directly or could not be reached.

# Algorithm of generating music melody based on single-exposure high dynamic range digital image using convolutional neural network

**Jiayue Cui[a] and Hongjun Wang[b,*]**
[a]Harbin Normal University, Harbin, China
[b]Harbin Conservatory of Music, Harbin, China

**Abstract.** Music is a very important art. However, professional music production has a relatively high investment cost and a single model, which cannot meet people's individual needs. Single-exposure high dynamic range (HDR) digital images based on convolutional neural networks (CNNs) can solve this problem. A comparative experiment was carried out. The final experimental results show that in terms of the quality of melody generation, the algorithm is almost the same as manual creation. In terms of generation speed, the CNN single-exposure HDR digital image model is 1.1 times faster than the traditional algorithm model, which proves its advantage of higher creative ability. Popularizing it in the music creation market at the current stage can effectively improve the efficiency of music creation and promote the intelligent development of the music market. © *2022 SPIE and IS&T* [DOI: 10.1117/1.JEI.31.5.051417]

**Keywords:** convolutional neural network; single-exposure high dynamic range digital image; music melody; algorithmic research.

Paper 220092SS received Feb. 15, 2022; accepted for publication May 6, 2022; published online Jun. 17, 2022; retracted Jun. 24, 2023.

## 1 Introduction

Since ancient times, composers have been creating perfect musical melodies in various ways and techniques to meet people's increasing demand for music. In recent years, science and technology such as artificial intelligence have continuously invaded the public's vision, and most of the industries that rely on scientific and intelligent technology have also ushered in new development opportunities and achieved transformation and upgrading, such as finance, logistics and transportation, processing and manufacturing and other value fields. On the other hand, scientific intelligence technology is gradually merging with the field of art.

As one of the representatives of the convolutional neural network (CNN), because of its powerful analysis and processing capabilities unmatched by other algorithms in image processing and can generate different images according to different data information.[1,2] Using it for the creation and generation of music melody can effectively create high-quality music melody, and can directly assist the composer's music creation, reduce the complex influencing factors in the creation process, improve the efficiency of music melody creation, and lower the threshold of creation.

This paper proposes a novel music melody algorithm based on single-exposure HDR digital image generation of CNN, which provides music creators with a melody creation platform based on intelligent science and technology. The algorithm and its application proposed in this paper can provide a new direction for the in-depth research of artificial intelligence, and can also provide a new idea for the intelligent research of music creation.

## 2 Related Work

In recent years, many scholars have conducted research on CNNs. Park et al. proposed an image generation method based on CNN, which can generate multiple images with different exposures

*Address all correspondence to Hongjun Wang, xieruifeng@hrbnu.edu.cn

1017-9909/2022/$28.00 © 2022 SPIE and IS&T

from a single input low dynamic range (LDR) image, thereby, improving high dynamic range (HDR) imaging. The proposed algorithm includes three steps, one is two-dimensional (2D) histogram estimation, the other is LDR image estimation based on neural network, and the third is to generate a set of optimal images with different exposures. This method first needs to generate image features by estimating a patch-based 2D histogram. The extracted features are used in the input layer of the neural network. Its function is to select a set of optimal LDR images, and then use the curvature-based contrast enhancement method generate a set of LDR images. The final experimental results show that the proposed method can use CNN technology to generate a set of optimal LDR images, and can improve HDR images.[3] In his article, Shi et al. showed a Comparative Genomic Hybridization pipeline based on a CNN that can synthesize realistic color three-dimensional (3D) holograms from a single RGB depth image in real time. He believes that the CNN algorithm has extremely high memory efficiency (<620 KB). It can not only run at a speed of 60 Hz and $1920 \times 1080$ pixels on a single consumer-grade graphics processing unit, but also can run interactively on mobile devices (iPhone 11 Pro at 1.1 Hz) and edge (Google Edge TPU at 2 Hz) devices. Finally, experiments prove that the CNN can generate 3D holograms with no spots, natural appearance and high resolution.[4] Wang et al. proposed a large-scale image annotation method based on CNNs. He believes that CNNs can improve the accuracy of original annotations to a certain extent, thereby enhancing the training effect of the model. So, he built a large-scale image annotation model MVIAACNN based on CNN. Finally, through experimental evaluation on MIRFlickr25K and NUS-WIDE datasets, and comparison with other methods, the effectiveness of MVAIACNN is proved.[5] Liu et al. proposed a new method based on CNN to solve the problem of unreliable information caused by instability in the digital forensics process. By adding a transformation layer, the obtained distinguishable frequency domain features are put into a conventional CNN model to identify template parameters of various types of spatial smoothing filtering operations, such as average, Gaussian, and median filtering. The experimental results on the composite database show that this method achieves better performance than some other applicable related methods, especially in the scene of small size and JPEG compression.[6] Guo et al. proposed a CNN-based HI construction method considering trend glitches. He first manipulates the learned features through convolution and pooling, and then constructs these learned features into a HI through a nonlinear mapping operation, and then uses anomalous area correction technology to detect and delete anomalous areas in HI. Different from the traditional method of manually constructing HI, his method aims to automatically construct HI, and finally he uses the bearing dataset to verify the effectiveness of the proposed method.[7] Sung-Pil proposes an extended CNN model to automatically extract protein-protein interaction information expressed in academic literature. The advantage of this model is to extend the feature-based CNN model designed for relation extraction. This model can also apply various global features to improve performance. In the experiment of the standard evaluation set AIMed for PPI extraction performance evaluation, the $F$-score experiment result is 78%, which proves that the performance of this model is 8.3% higher than the world's best performance obtained so far. In addition, the experimental results also show that the CNN model exhibits high performance in extracting protein-protein interactions, and does not require complex language processing for feature extraction.[8] In summary, after recent years of exploration, CNNs have been applied to various fields, but there is not much research in music melody generation, and more in-depth exploration is needed.

## 3 Algorithm for Generating Music Melody from Single-Exposure HDR Digital Image Based on Convolutional Neural Network

### 3.1 *Convolutional Neural Network*

The brain is composed of many neurons.[9] If a single perceptron is equivalent to a neuron, the combination of multiple perceptrons can form a neural network structure, as shown in Fig. 1.

When there are two neurons in the output layer, different classification problems can also be dealt with at the same time,[10] as shown in Fig. 2.

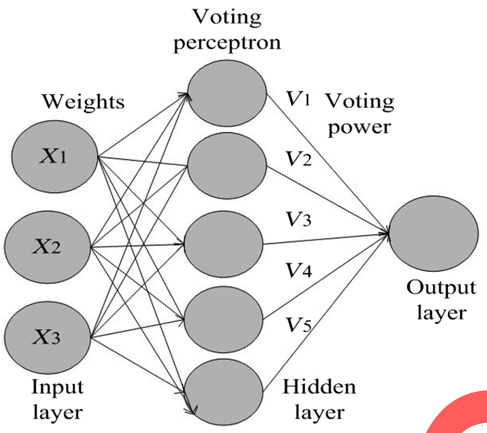

**Fig. 1** Neural network structure.

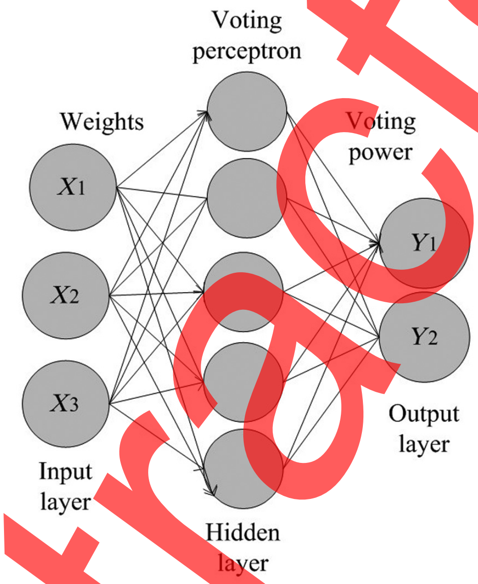

**Fig. 2** Multi-class network structure.

The CNN belongs to a multi-layer network structure, which is mainly composed of a convolutional layer, a pooling layer, and a fully connected layer,[11] as shown in Fig. 3.

In the convolution kernel, the neighborhood range feature corresponding to the larger weight value contributes more to the final result.[12] The convolution kernel is equivalent to a weight template. It slides and walks in the image matrix. After sliding once, a convolution calculation is performed, and the result is used as the response of the corresponding pixel on the image. As shown in Fig. 4, assuming that the Input Figure is a matrix with a size of and a convolution

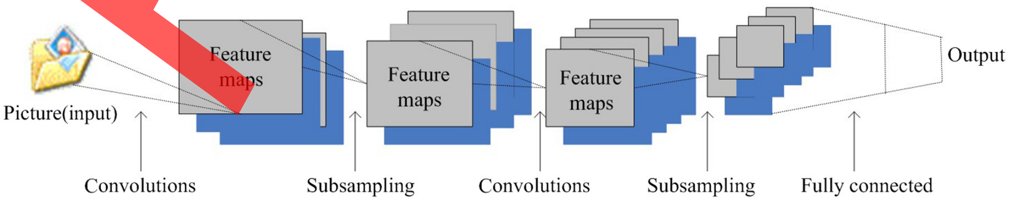

**Fig. 3** CNN structure.

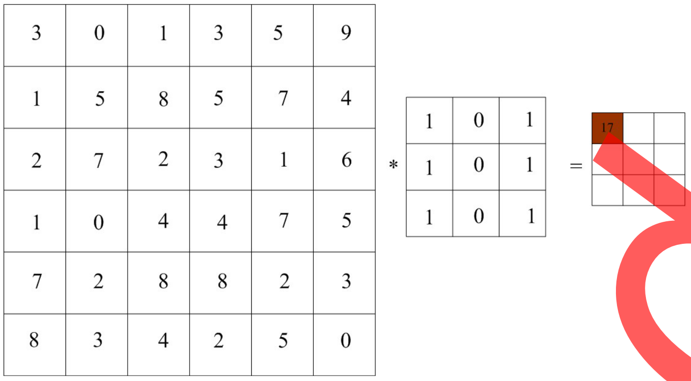

**Fig. 4** Convolution calculation process.

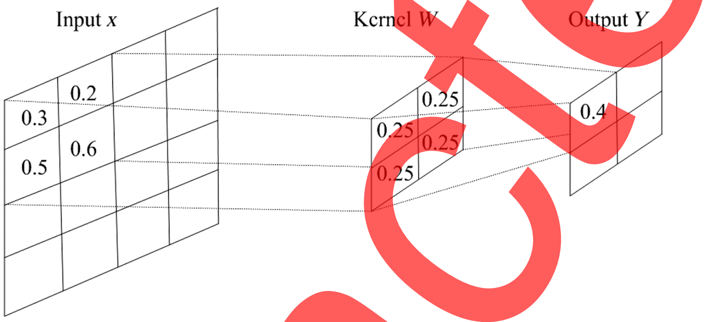

**Fig. 5** Mean pooling process.

kernel, the sliding step is set to 1, and the red area in the image is calculated by the convolution kernel to obtain the value of the (0,0) position as 17.

The matrix obtained after convolution calculation also needs dimensionality reduction processing.[13,14] The pooling layer can reduce the resolution of the image, can reduce the matrix size, and can increase the calculation rate. Commonly used pooling methods include mean pooling and maximum pooling, as shown in Fig. 5. This is the mean pooling process, where the weight matrix values are all 0.5 and the sliding step size is 2.

Figure 4 shows the maximum pooling. In the weight matrix, there is only one 1, and the rest are 0. The position corresponding to one is the maximum point of the coverage area of the input image, and the sliding step is also 2.[15,16] It can be seen from Figs. 5 and 6 that the images after the pooling operation are all reduced to a quarter of the original image.

In the CNN, the fully connected layer can map the latent features learned by the convolutional layer and the pooling layer to the labeled sample space to realize the image classification

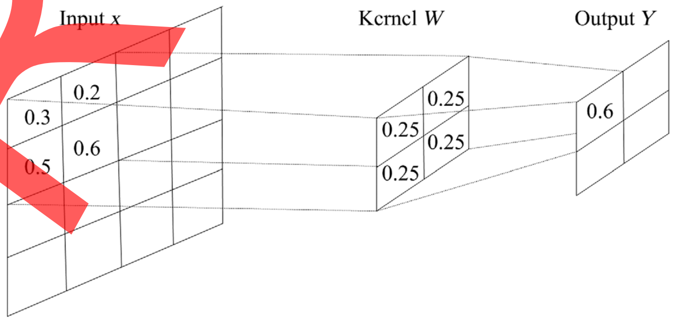

**Fig. 6** Maximum pooling process.

function. The idea of convolution can also be used, that is, the convolution calculation is performed through the set convolution kernel.

## 3.2 *Single-Exposure HDR Digital Image Generation*

### 3.2.1 *Overview of HDR digital image*

HDR digital images are different from ordinary images in that it has a larger dynamic range and richer detailed information. Although it can be displayed on ordinary monitors by tone mapping, the difference information of many HDR images will be weakened after compression, making it indistinguishable to the naked eye. Therefore, algorithms are generally used to achieve tone mapping display in research. The algorithm is a global tone mapping algorithm, which can linearly scale the brightness in the image twice. The first linear scaling is carried out by logarithmic average brightness and proportional scaling, which is expressed as[17]

$$L' = \frac{\alpha}{L_{\text{avg}}} L_w, \tag{1}$$

$$L_{\text{avg}} = \exp\left(\frac{1}{N}\sum \log(L_w + \delta)\right). \tag{2}$$

Then, performing a second linear compression transformation on the brightness value after linear scaling so that the highest brightness value will not be mapped to 1. The expression is as follows:

$$L_{\text{out}} = \frac{L'}{1 + L'}. \tag{3}$$

At the same time, the introduction of $L_{\text{white}}$ expands the above equation into a controllable function, the formula is as follows[18]:

$$L_{\text{out}} = \frac{L'\left(1 + \frac{L'}{L_{\text{white}}^2}\right)}{1 + L'}. \tag{4}$$

Because subjective visual perception may be affected by various factors, different people have different visual perceptions, we need to add some objective evaluation indicators for comparative analysis.

The peak signal-to-noise ratio is an engineering term that expresses the ratio of the maximum possible power of a signal to the destructive noise power that affects its representation accuracy. It can be defined by a simple mean square error, expressed as[19]

$$\text{PSNR} = 10 \cdot \log_{10}\left(\frac{\text{MAX\_}I^2}{\text{MSE}}\right). \tag{5}$$

In Eq. (5), the maximum value of the test image pixel is the mean square error between the test image and the reference image, and the calculation formula is as follows[20]:

$$\text{MSE} = \frac{1}{\text{MN}}\sum\sum \left| |I_i - I_i'| \right|^2. \tag{6}$$

The theoretical basis of SSIM is the self-adaptive adjustment of the structure in the scene of the human visual system. By comparing the changes of the image structure information to judge the similarity of the images, so as to obtain an objective quality evaluation.

It measures the composite effect of image brightness, contrast and structural changes. The evaluation model is expressed as

**Table 1** The meaning of each parameter of the formula.

| Parameter | Implication |
|---|---|
| $u_x$ | Test image mean |
| $u_y$ | Standard image mean |
| $\sigma_x$ | Standard deviation |
| $\sigma_y$ | Standard deviation |
| $\sigma_{xy}$ | Covariance |
| $C_1$ | Small constant |
| $C_2$ | Small constant |
| $C_3$ | Small constant |

$$\mathrm{SSIM}(x,y)=l(x,y)^{\alpha}c(x,y)^{\beta}s(x,y)^{\gamma}. \tag{7}$$

In Eq. (7), the three components of $l$, $c$, $s$ are calculated by the following formulas as[21]

$$l(x,y)=\frac{2u_xu_y+C_1}{u_x^2+u_y^2+C_1}, \tag{8}$$

$$c(x,y)=\frac{2\sigma_x\sigma_y+C_2}{\sigma_x^2+\sigma_y^2+C_2}, \tag{9}$$

$$s(x,y)=\frac{2\sigma_{xy}+C_3}{\sigma_x\sigma_y+C_3}. \tag{10}$$

The meaning of each parameter in Eqs. (8) to (10) is shown in Table 1.

The parameters $\alpha$, $\beta$, and $\gamma$ can be used to adjust the proportions of the three components in the model. Generally, the three components are equally important by default, namely, $\alpha=\beta=\gamma=1$. When $C_3=C_2/2$ is set, it can be simplified to the formula as follows[22]:

$$\mathrm{SSIM}(x,y)=\frac{(2\mu_x\mu_y+C_1)(2\sigma_{xy}+C_2)}{(\mu_x^2+\mu_y^2+C_1)(\sigma_x^2+\sigma_y^2+C_2)}. \tag{11}$$

The result of function calculation is between [0,1]. The more similar the two images, the closer the calculation result is to one.

### 3.2.2 *Single-exposure HDR digital image generation model*

*Global expansion model.* Global expansion refers to the use of the same function to process all the pixels of the image. This kind of method is usually simple to calculate, and is usually realized by a simple curve function. It performs better for high-quality HDR images, but may experience loss of detail in poor-quality images.

To obtain suitable HDR images, a simple extension function is proposed, the formula expression is as follows[23]:

$$L'=k\left(\frac{L-L_{\min}}{L_{\max}-L_{\min}}\right)^{r}. \tag{12}$$

The meaning of each parameter of Eq. (12) is shown in Table 2.

**Table 2** The meaning of each parameter of the formula.

| Parameter | Implication |
|---|---|
| $L$ | Image original brightness value |
| $L_{\text{min}}$ | Image brightness minimum |
| $L_{\text{max}}$ | Image brightness maximum |
| $k$ | The maximum brightness value that the HDR display can display |
| $\gamma$ | Non-linear characteristics of the control curve |

At that time, it was linear expansion, and it was adopted as 1 during the experiment. The result shows that when the average brightness is the same, the expanded HDR image classification expansion model looks better.

Through the improvement of the algorithm, a curve-based adaptive global expansion function is proposed. First, the brightness value of the image is calculated by the brightness formula, the expression is as follows:

$$L = 0.2126R + 0.7152G + 0.0722B. \tag{13}$$

Then calculating the geometric mean of the brightness, the formula is as follows[24]:

$$L_{\text{avg}} = \exp\left(\frac{1}{N}\sum_{i=1}^{N} \log(L(i) + \varepsilon)\right). \tag{14}$$

In Eq. (14), it is the number of pixel values, which is a very small positive number, used to prevent the influence of singular pixel values, and then using the geometric mean of brightness to calculate a key value. The calculation formula is as follows:

$$k = \frac{\log L_{\text{avg}} - \log L_{\text{min}}}{\log L_{\text{max}} - \log L_{\text{min}}}. \tag{15}$$

In Eq. (15), $L_{\text{max}}$ and $L_{\text{min}}$ are the maximum and minimum values of image brightness, respectively, and then calculating the overexposure area $P_{ov}$ whose brightness value is >254, and then combining the geometric mean value of brightness $L_{\text{avg}}$ and the key value $k$ to obtain the final $\gamma$, the calculation formula is as follows:

$$\gamma = 2.4379 + 0.2319 \log L_{\text{avg}} - 1.1228k + 0.0085P_{\text{ov}}. \tag{16}$$

Finally, the expanded brightness is calculated by Eq. (12), and then three channels of $R$, $G$, and $B$ are restored to obtain the final HDR image.

*Classification expansion model.* Classification expansion refers to dividing different contents of LDR into different areas, and then different areas are expanded using different methods. Usually such methods are classified according to different exposure levels.

The classification expansion function formula is as follows[25]:

$$f(I(p)) = \begin{cases} S_1 \cdot I(P) & I(P) \leq \omega \\ S_1 \cdot \omega + S_2 \cdot (I(P) - \omega) & I(P) > \omega \end{cases}. \tag{17}$$

Among it:

$$S_1 = \frac{\rho}{\omega} \quad S_2 = \frac{1-\rho}{I(P)_{\max} - \omega}. \tag{18}$$

In Eq. (17), $I(P)$ is the normalized brightness value, and the maximum value is 1.

The research also proposes an extended model based on the scene classifier. The scene is divided into three brightness levels by a classifier, and then the three brightness levels and combined parameters are used to map the image to a range that matches the dynamic range of the scene. The calculation formula is as follows:

$$L' = \frac{L_d L_{\max}}{\rho(I - L_d) + L_d}. \tag{19}$$

In the formula, $L'$ represents the brightness of the image after expansion, $L_d$ represents the brightness of the original LDR image, the maximum brightness of the $L_{\max}$ image after expansion, and $\rho$ is the minimum value that is not a black pixel.

*Extended mapping model.* Extended mapping refers to the inverse tone mapping transformation of LDR to extend the dynamic range, and then different methods are used to restore the details of the overexposed area. But this kind of method does not have obvious effect when the overexposure area of the image is large.

In the research, a global inverse tone mapping operator is proposed, which is obtained by inverting the tone mapping operator. This operator has fewer parameters and can easily control the extended range of the image. The formula is as follows:

$$L_w(x) = \frac{1}{2} L_{w,\max} L_{\text{white}} \cdot \left( L_d(x) - 1 + \sqrt{\left(1 - L_d(x)^2 + \frac{4}{L_{\text{white}}^2} L_d(x)\right)}\right). \tag{20}$$

In Eq. (20), $L_{w,\max}$ is the maximum output brightness of the expanded image, and $L_{\text{white}}$ is a parameter that can determine the shape of the expanded curve, which is proportional to the contrast.

### 3.3 *Melody Mixing Rules*

In music creation, mostly mixed sounds are used, and a single melody is basically not used or rarely used. A single melody can be popularly understood as the inconsistency between the internal part of the melody and the entire melody. To make the generated music more aesthetic, it is necessary to follow the creation rules, as shown in Table 3:

**Table 3** Music melody mixing rules.

| Component | Scale range |
|---|---|
| Within bars | [0,2, 3,4] |
| Between bars | [0,2, 3,4, 5,7] |
| Push to the interval sample value corresponding to the climax | [7,9, 10,12] |
| The number of second intervals ($N_{i=1}$) | 1 ($N_{i=1} \le 2$) |
| | 0 ($N_{i=1} > 2$) |

**Table 4** Experimental parameter initialization setting.

| Parameter | Value |
|---|---|
| Global learning rate $(\varepsilon)$ | 0.001 |
| Initial parameter value $(\theta)$ | 0.9 |
| Numerical stability $(\delta)$ | 108 |
| Decay rate $(\rho)$ | 0.0 |
| Neuron disconnection rate | 0.3 |
| Number of iterations | 20 times |

## 4 Generating Music Melody from Single-Exposure HDR Digital Image Based on Convolutional Neural Network

### 4.1 *Test Data and Parameter Settings*

The data collected in this article are 120 classical music scores of 3/4 beats and 70 to 180 beats per minute, and then the scores are converted into audio format by professional conversion technology, and these audios are edited into a duration of 3 s. A total of 9815 pieces of unit audio have been edited. In the CNN single-exposure HDR digital image model, after repeating experiments, the initial settings of the experimental parameters are shown in Table 4:

#### 4.1.1 *Generate test*

This test uses the turing test. Ten melodies were selected for the test, five of which were generated from single-exposure HDR digital images of CNNs, and five were created by composers in the dataset music library. Ten testers were randomly invited to participate in the experimental test. The melody generated by the single-exposure HDR digital image based on the CNN and the melody created by the composer are played alternately to the tester. The sequence is shown in Table 5. The playback interface only has the track number, and other sample information is not displayed. After playing, asking them to rate the melody beat quality, melody logic, and sample satisfaction based on their personal feelings. Scores range from 0 to 100, with 0 being very bad and 100 being very good. Because the testers' music theory knowledge and subjective

**Table 5** Music playback order.

| Serial number | Music name | Author |
|---|---|---|
| 1 | Song of Spring | Artist |
| 2 | Generate1 | Algorithm generation |
| 3 | Generate2 | Algorithm generation |
| 4 | Variations on the little star | Artist |
| 5 | Generate3 | Algorithm generation |
| 6 | Generate4 | Algorithm generation |
| 7 | Moonlight variations | Artist |
| 8 | Generate5 | Algorithm generation |
| 9 | June boat song | Artist |
| 10 | Dream song | Artist |

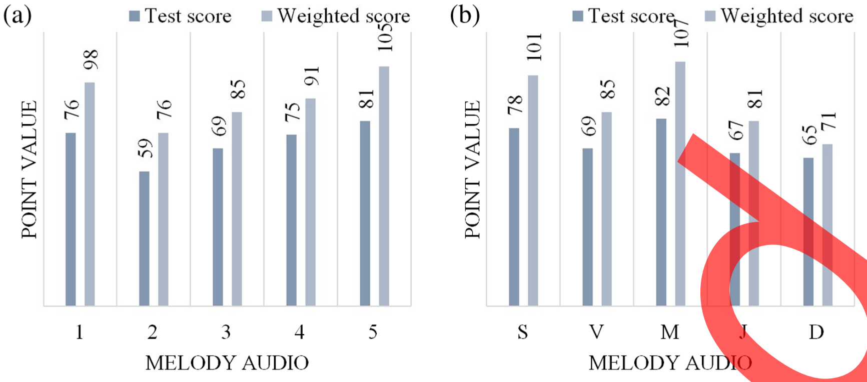

**Fig. 7** Test score statistics. (a) Beat quality score and weighted score of melodies generated by CNN single-exposure HDR digital images; (b) beat quality scores and weighted scores for melodies composed by the composer.

preferences have individual differences, this paper will again carry out weighted statistics on the testers' scores, and the score statistics results are shown in Figs. 7, 8, and 9.

Figure 7(a) shows the beat quality score and weighted score of the melody generated by the CNN single-exposure HDR digital image.

Figure 7(b) shows the beat quality score and weighted score of the melody created by the composer.

It can be seen from Fig. 7 that the average test score of the melody generated by the CNN single-exposure HDR digital image is 72 points, and the average weighted score is 91 points. The average test score of the composer's melody creation is 72.2 points, and the average weighted score is 89 points.

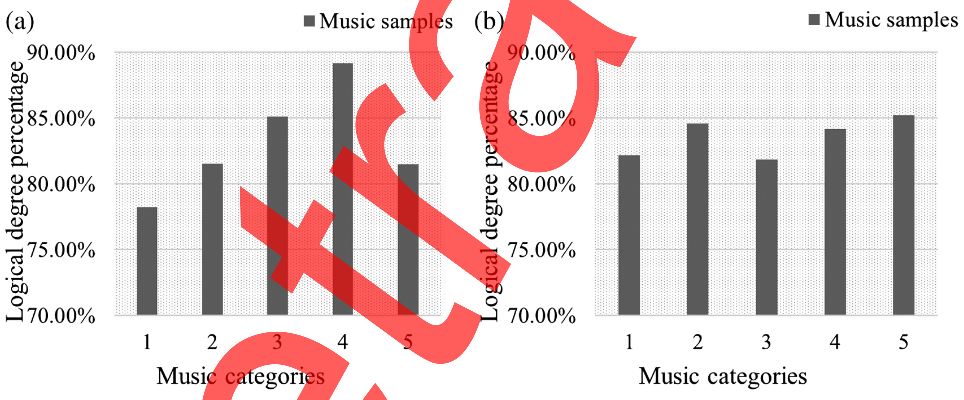

**Fig. 8** Statistic results of music melody logic degree.

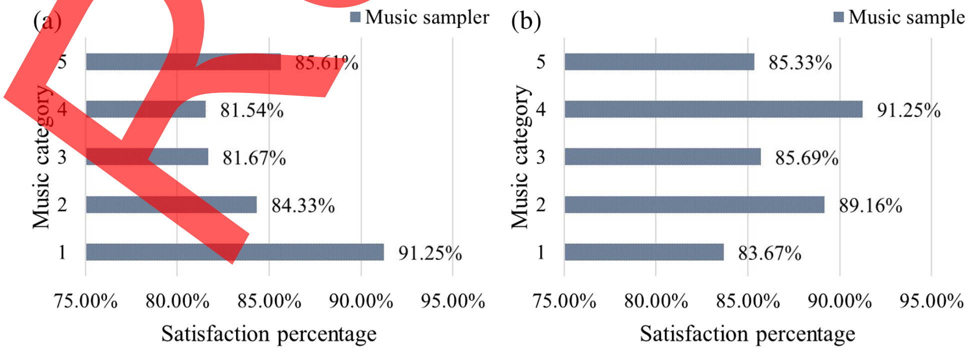

**Fig. 9** Satisfaction statistical results of music samples.

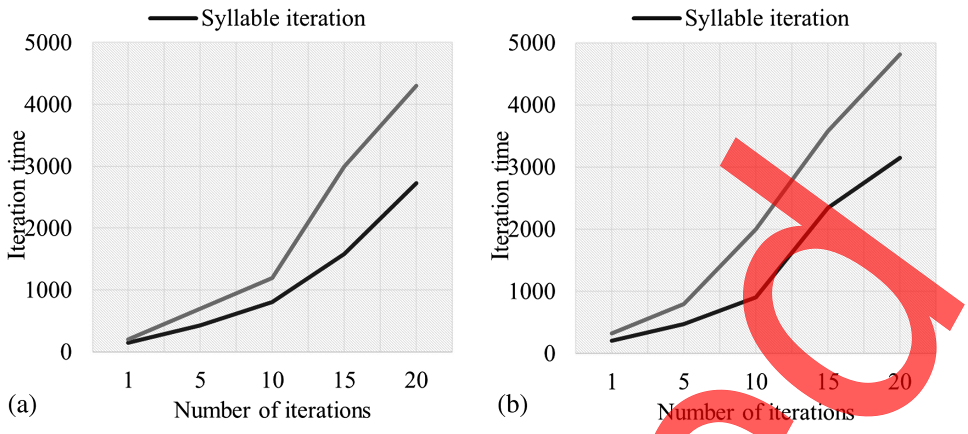

**Fig. 10** Iteration time test results.

Figure 8(a) shows the logic degree of the melody generated by the CNN single-exposure HDR digital image.

Figure 8(b) shows the logic of the composer's creation of the melody.

It can be seen from Fig. 8 that the average logical degree of the melody generated by the CNN single-exposure HDR digital image is about 83.11%. The average logical degree of the melody created by the composer is 83.60%.

Figure 9(a) shows the sample satisfaction degree of the melody generated by the CNN single-exposure HDR digital image.

Figure 9(b) shows the sample satisfaction rate of the composer's creation of the melody.

It can be seen from Fig. 9 that the average satisfaction rate of samples generated by the CNN single-exposure HDR digital image of the melody is 84.88%. The average satisfaction rate of samples generated by the composer's melody is 87.02%.

### 4.2 *Comparison Test of Melody Generation Time*

To maintain the objectivity of the experiment, the melody generation time comparison test and the melody generation test use exactly the same data. There are a total of 120 music scores in 3/4 beats, and the playing speed is 70 to 180 beats per minute. The music scores are converted into audio formats through professional conversion technology, and these audios are edited into unit audio with a duration of 3 s. A total of 9815 pieces are edited. A clipped unit of audio consists of 1 to 3 bars. The number of model iterations is set to 20. The test results are shown in Fig. 10:

Figure 10(a) shows the iteration time of single-exposure HDR digital image based on CNN.

Figure 10(b) shows the iteration time of the traditional algorithm.

It can be seen from Fig. 10 that after 20 iterations, it takes 4300 s to generate the melody based on the CNN single-exposure HDR digital image iteration, of which the syllable iteration time is 2730 s. The traditional algorithm iteratively takes 4820 s to generate the melody. The syllable iteration time is 3150 s.

## 5 Discussion

By comparing the experimental data of the melody generated by the single-exposure HDR digital image model based on the CNN with the composer's melody's comparison test data, the following conclusions can be drawn:

1. The beat quality scores of the single-exposure HDR digital image model generated by the CNN and the human work music melody are relatively evenly distributed, indicating that the quality of the melody generated by the model and the quality of the melody created by humans are almost the same. The experiment also found that the testers could not clearly distinguish between music created by humans and music created by algorithms. This also

shows that the CNN single-exposure HDR digital image model has a better effect in music connection processing.

2. The single-exposure HDR digital image model based on the CNN takes more time to generate the melody during or after the syllable iteration than traditional algorithms. It shows that the melody generated by the CNN single-exposure HDR digital image model can play its own advantages in commercial development at a lower time cost.

## 6 Conclusion

Using artificial intelligence to study efficient music generation algorithms for music creation can not only enrich people's spiritual world, but also promote the intelligent development of the music industry. The music melody algorithm represented by the CNN single-exposure HDR digital image model can create high-quality melody on the basis of following theories, and deliver the most intuitive and beautiful auditory experience to the audience. In addition, it can generate music melody in a relatively short time, so it also has a good development prospect in commercial promotion.

In this paper, it is a good attempt to verify the practicability and effectiveness of the CNN single-exposure HDR digital image generation algorithm for music melody by controlled experiments, but there are still many shortcomings. The depth and breadth of the research in this paper are not enough. In future research work, we will further study how to construct a multi-style music generation algorithm from a single-exposure HDR digital image of a CNN, and to experiment and improve the algorithm depth, and improve the synthesis of music. The quality of music is further researched and explored.

## Acknowledgments

This work was supported by the National Social Science Fund Art Project under project No. 2020BD03601; project category: normal project; authorization No. 20BD053; and project name: the production and development of Russian vocal art in China in 20th. The authors declare no conflict of interest. This work is completed by Jiayue Cui and Hongjun Wang, Jiayue Cui is responsible for writing- editing data curation, writing- original draft preparation, Hongjun Wang is responsible for resources; supervision; data analysis, visualization, investigation.

## References

1. H Wei and N. Kehtarnavaz, "Determining number of speakers from single microphone speech signals by multi-label convolutional neural network," in *IEEE IECON*, IEEE (2018).
2. X. Gong et al., "ChMusic: a traditional Chinese music dataset for evaluation of instrument recognition," (2021).
3. K. Park et al., "An optimal low dynamic range image generation method using a neural network," *IEEE Trans. Consum. Electron.* **64**(1), 69–76 (2018).
4. L. Shi et al., "Towards real-time photorealistic 3D holography with deep neural networks," *Nature* **591**(7849), 234–239 (2021).
5. R. Wang et al., Large scale automatic image annotation based on convolutional neural network," *J. Vis. Commun. Image Represent.* **49**(Nov.), 213–224 (2017).
6. A. Liu et al., Smooth filtering identification based on convolutional neural networks," *Multimedia Tools. Appl.* **78**(19), 26851–26865 (2019).
7. L. Guo et al., "Machinery health indicator construction based on convolutional neural networks considering trend burr," *Neurocomputing* **292**(May 31), 142–150 (2018).
8. C. Sung-Pil, "Extraction of protein-protein interactions based on convolutional neural network (CNN)," *KIISE Trans. Comput. Pract.* **23**(3), 194–198 (2017).
9. C. H. Liu and C. K. Ting, "Computational intelligence in music composition: a survey," *IEEE Trans. Emerg. Top. Comput. Intell.* **1**(1), 2–15 (2017).
10. X. Zheng et al., "Algorithm composition of Chinese folk music based on swarm intelligence," *Int. J. Comput. Sci. Math.* **8**(5), 437–446 (2017).

11. Z. Lv et al., "Deep learning for security in digital twins of cooperative intelligent transportation systems," *IEEE Trans. Intell. Transport. Syst.* (2021).
12. W. Zhang et al., "Melody extraction from polyphonic music using particle filter and dynamic programming," *IEEE/ACM Trans. Audio Speech Lang. Process.* **26**(9), 1620–1632 (2018).
13. S. A. Herff, R. T. Dean, and K. N. Olsen, "Interrater agreement in memory for melody as a measure of listeners' similarity in music perception," *Psychomusicol.: Music Mind Brain* **27**(4), 297–311 (2017).
14. D. K. Jain, X. Lan, and R. Manikandan, "Fusion of iris and sclera using phase intensive rubbersheet mutual exclusion for periocular recognition," *Image Vision Comput.* **103**, 104024 (2020)
15. C. K. Ting, C. L. Wu, and C. H. Liu, "A novel automatic composition system using evolutionary algorithm and phrase imitation," *IEEE Syst. J.* **11**(3), 1284–1295 (2017).
16. S. Ding et al., "Stimulus-driven and concept-driven analysis for image caption generation," *Neurocomputing* **398**, 520–530 (2019).
17. J. Abeer and G. Schuller, "Instrument-centered music transcription of solo bass guitar recordings," *IEEE/ACM Trans. Audio Speech Lang. Process.* **25**(9), 1437–1446 (2017).
18. R. T. Schirrmeister et al., "Deep learning with convolutional neural networks for decoding and visualization of EEG pathology," *Human Brain Mapp.* **38**(11), 5391–5420 (2017).
19. L. Jae-Hong et al., "Diagnosis and prediction of periodontally compromised teeth using a deep learning-based convolutional neural network algorithm," *J. Periodontal Implant Sci.* **48**(2), 114–123 (2018).
20. X. Ma et al., "Deep learning convolutional neural networks for pharmaceutical tablet defect detection," *Microsc. Microanal.* **26**(S2), f1–f88 (2020).
21. S. A. Azer, "Deep learning with convolutional neural networks for identification of liver masses and hepatocellular carcinoma: a systematic review," *J. Gastrointest. Oncol.* **11**(12), 1218–1230 (2019).
22. H. A. Haenssle et al., "Reply to the letter to the editor 'Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists' by H. A. Haenssle et al.," *Ann. Oncol.* **30**(5), 854–857 (2019).
23. D. Herremans and C. H. Chuan, "The emergence of deep learning: new opportunities for music and audio technologies," *Neural Comput. Appl.* **32**, 913–914 (2020).
24. K. Yasaka et al., "Deep learning with convolutional neural network in radiology," *Jpn. J. Radiol.* **36**(4), 257–272 (2018).
25. X. Ma et al., "Application of deep learning convolutional neural networks for internal tablet defect detection: high accuracy, throughput, and adaptability," *J. Pharm. Sci.* **109**(4), 1547–1557 (2020).

**Jiayue Cui** holds a doctoral degree in musicology and dancology from Harbin Conservatory of Music, China. Currently, she is an associate professor at the College of Master, Harbin Normal University. She has written three academic books and published more than 10 papers in national and provincial core journals, and she is the first participant of the 2020 National Social Science, Art and Science Planning Project. She has completed 11 provincial-level scientific research projects and won the second prize of Heilongjiang Provincial Excellent Achievements in Art and Science Planning.

**Hongjun Wang** received the mastery of professor, the tutor of doctor, bass singer, the deputy director of 1st academic committee of Harbin Conservatory of Music, and the deputy director of Vocal Opera Department. He is the fourth member of the CPPCC, the national social artificial science planning project essayist, director of Chinese Music Oral History Research Association, member of Heilongjiang Musician Association, and professor of Lingnan Normal University.