ingegneria sismica
International Journal of Earthquake Engineering
Anno XLIII - Num. 1 - 2026

Choral Works Style Transfer and Voice Part Balance Algorithm based on Generative adversarial networks

Zhipeng Li[1,*]

[1] School of Music and Dance, Lanzhou University of Arts and Science, Lanzhou, 730000, Gansu, China

SUMMARY: *With the continuous development of artificial intelligence technology and deep learning methods in the field of music information processing, style transfer and multi-voice balance control of choral works have gradually become an important issue in computational musicology. This paper aims to construct a style transfer and voice part balance algorithm for choral works based on generative adversarial networks. Based on 312 choral works and 6840 valid sample fragments, this study jointly collects audio and music data, and completes the unified preprocessing through timing alignment, loudness normalization and multi-modal feature coding. Focusing on the characteristics of chorus style expression and voice organization, we extract information such as spectrum envelope, harmony density, rhythm intensity and voice energy ratio, and introduce cycle consistency constraint, content preservation constraint and voice balance regularization term into the bidirectional generative adversarial framework to realize the collaborative modeling of chorus style transfer and hierarchical optimization. Experimental results show that the proposed method achieves 0.87, 91.6%, 0.041 and 4.36 in style similarity, main melody retention rate, voice part balance error and subjective listening score, respectively, which are better than basic CycleGAN, Transformer-GAN and AutoMix-Net. The results show that the algorithm has good application value in chorus intelligent composition and digital music processing.*

KEYWORDS: *Generative adversarial networks; Choral style transfer; Voice balance; Computational musicology*

1 Introduction

Choral works are both the result of the organization of the musical structure and the concentrated presentation of the collective sound aesthetic. Compared with solo or instrumental fragments, choral texts often contain multiple information such as melody direction, harmonic support, texture density, voice part distribution and spatial level. The differences in these dimensions of different styles of works are not only reflected by "different timbres" or "different speeds", but also by the overall changes in the master-slave relationship between voice parts, spectrum energy allocation and syntactic advancement. Because of this, choral style analysis has not only been the focus of traditional musicology, but also gradually become an important research object in the fields of computational musicology, music information retrieval and intelligent generation. With the continuous advancement of deep learning in audio modeling, singing voice separation, symbolic music generation and automatic mixing, researchers have begun to try to use learnable models to describe complex

*hotzaradr@163.com
https://doi.org/10.65102/is2026310 Published: 30 April, 2026

music styles, and realize style mapping, structure preservation and auditory quality control in the generation process [1].

In the existing research, music style transfer mostly focuses on solo timbral control, melody generation or single-track symbol sequence, and the research on the complex object of multi-part chorus is still relatively limited [2]. The reason is that a choral work is not a simple superposition of a number of single lines. Once the style transfer is separated from the cooperative relationship of voice parts, the model may generate similar results on the local timbral or surface texture, but it is prone to the phenomenon that the main melody is submerged, the low voice part is insufficient support, the middle voice part is stacked and cloudy, or the high frequency region is abnormally prominent. In other words, choral style transfer is not just "writing one kind of music like another kind of music", but integrating style feature transfer and voice part balance control into a unified computing framework while maintaining the basic content of the work to be recognizable. The traditional rule-driven method is difficult to deal with the coupling relationship between multi-scale style features, and the generative model only relying on reconstruction error is difficult to ensure the stability of energy distribution between voices.

Generative adversarial networks provide a feasible path for this task. Through the adversarial game between the generator and the discriminator, the generated results are pushed to constantly approach the distribution of the target domain, which has shown strong distribution fitting ability in tasks such as image style transfer, voice conversion and music generation. If it is introduced into the chorus scene, the model can not only learn the statistical differences in spectral texture, rhythm density and harmonic organization of different styles of works, but also combine the score prior and multi-part constraints to preserve the melody skeleton, syntactic boundaries and longitudinal harmonic structure during the transfer process. However, the existing related researches focus more on singing voice separation, automatic mixing or conditional music generation, and rarely deal with the dual-objective synergy problem of "style transfer and voice part balance" [3]. This means that specialized modeling for data representation, feature extraction, and loss design is still required to obtain truly usable results for choral composition assistance, digital music production, and instructional analysis.

1.1 Problem Formulation

In the process of traditional chorus adaptation and style reconstruction, the creators usually rely on experience judgment to complete the adjustment of voice texture, timbrel imagination and style closeness. This method is not invalid, but the efficiency is restricted by personal experience, and the results of style transfer are not easy to quantify and compare. When the scale of the work increases, the structure of the voice part is complex or the target style is different, it is often difficult to balance the transfer effect, content maintenance and voice part only by manual trial writing and repeated listening. Although some existing music generation methods can simulate specific styles on local segments, they often compress the relationship of multiple voice parts into a unified representation, which weakens the functional boundaries of female high, female low, male high, male low and other voice parts. Other automatic mixing methods can optimize the loudness distribution, but do not really understand the syntactic organization and harmonic direction within the chorus style. Although the generated results may be acceptable in general music, they may not meet the requirements of the hierarchical organization of choral genres, let alone stably present the characteristics of choral writing in the target style.

Based on this, the core issue of this paper is not the timbre replacement of a single vocal track, nor the automatic balancing in the ordinary sense, but how to construct a generative

adversarial network model for multi-voice choral works, so that it can learn the mapping relationship between the source style and the target style under the condition of unpaired or weak paired samples. In the generation process, the voice energy proportion, frequency domain occupation and melody saliency are synchronously constrained, so as to obtain the output result with both style consistency and auditory balance. Aiming at this goal, we jointly encode multi-voice audio features, musical notation features and voice balance indicators, and introduce balance optimization constraints in addition to style discrimination to alleviate the problems of voice masking, spectrum congestion and hierarchical imbalance after migration.

1.2 Research Contribution

This paper aims to construct a style transfer and voice part balance optimization method for choral works based on generative adversarial networks, so that style mapping no longer stays at the surface timsound simulation, but is implemented into the cooperative organization of multiple voices, harmony texture maintenance and balance control.

This paper establishes a sample set around mixed choral works, and jointly collects multi-voice audio, corresponding music score and voice part annotation information to form multi-modal training data that can be used for style learning, balance modeling and comparative verification.

In the preprocessing stage, this paper performs segmentation, alignment, loudness normalization and time-frequency transformation on chorus recordings from different sources, and encodes the pitch, duration, beat and voice part trajectory of the score data to ensure the consistency of the model input in time scale and structural semantics.

At the feature modeling level, we extract the spectrum envelope, rhythm density, harmony progression and texture distribution features that reflect the differences in chorus styles, and construct the energy ratio, frequency band coverage, melody prominence and masking indicators that describe the balance state of voice parts, so that the style transfer and balance evaluation can be carried out in a unified feature space.

At the level of algorithm design, this paper introduces a bidirectional generative adversarial network framework to complete the style domain mapping, and combines multi-scale convolution discrimination, cycle consistency constraint, content preservation constraint and voice part balance regularization term for co-training to improve the model's ability to generate complex chorus textures and maintain key voice part relationships.

The experimental results verify the effectiveness of the model from multiple perspectives such as style similarity, voice part balance error, main melody retention rate and listening perception evaluation. The adaptability and stability of the proposed method under different voice part configurations are tested by comparing with the traditional style transfer method, the common generation model and the variant without balance constraints.

The rest of this paper is arranged as follows: Section 2 reviews related research; Section 3 introduces data construction, feature extraction, and generative adversarial networks algorithms; Section 4 gives experimental results and discussions; Section 5 concludes the paper and explains future research directions.

2 Related Research

In recent years, the research on intelligent analysis and generation of multi-voice music has gradually shifted from single sound source processing to structural modeling. Chen et al. improved the chorus separation task by constructing synthetic data with performance differences, so that the model could more fully learn the spectrum distribution law under the

condition of multiple singing overlaps [4]. This study proves that the augmented data strategy has a positive effect on complex chorus scenes, but its focus is still on improving the accuracy of sound source separation, and how to maintain the voice part relationship in the process of style transfer is not carried out. Jeon et al. released a multi-singing separation evaluation dataset, which provides a unified benchmark for the objective comparison of complex singing textures [5]. This dataset improves the reproducibility of research, but mainly serves the separation task, and the support for chorus style representation and balance control is still limited. Yu et al. and Richard et al. respectively promoted the decoupling modeling of singing voice from the perspectives of diffusion model and differentiable unsupervised separation. The results show that the deep generative model has the ability to deal with mixed signals of multi-person singing, but such methods pay more attention to "split voice" and have not really solved the problem of "how to maintain the overall level of chorus after style transfer" [6, 7].

In the study of automatic mixing, Martinez-Ramirez et al. used deep learning and cross-domain data to realize automatic music mixing, indicating that the model can learn the potential rules between loudness, frequency band and spatial allocation [8]. Koszewski et al. further established an automatic mixing system based on one-dimensional Wave-U-Net autoencoder to achieve stable balance regulation at the signal level [9]. Wu and Horner introduced the diffusion model to deal with music mixing tasks, which improved the detail control ability in complex scenes [10]. These studies show that machine learning methods have been able to automatically optimize the energy ratio of multi-track audio. However, the goal of automatic mixing is usually to obtain a general sense of clarity and balance, which does not directly correspond to the configuration of chorus parts in a specific style context, so it is difficult to directly transfer the results to the composite task of "style transfer and voice part reorganization".

In the field of music style transfer and conditional generation, Mukherjee and Mulimani discussed the problem of music composition with style transfer constraints, indicating that style can be embedded into the generation process as an independent condition [11]. Zhang et al. proposed StyleSinger to achieve style control in cross-domain singing synthesis. Zhang et al. further proposed a zero-shot singing style transfer framework with multi-level control ability, which enhanced the model's ability to adjust the details of singing expression [12, 13]. Huang et al. used diffusion model for timbral style transfer, and Hong et al. jointly modeled voice and accompaniment in text-to-song generation, which promoted the development of controllable music generation [14, 15]. On the other hand, Sulun et al., Wang et al., Zhang et al., and von Rutte et al., have enriched the conditional representation space of music generation models from the perspectives of emotional conditions, style conditions, harmony perception and controllable feature coding [16-19]. Related results show that current research has been able to learn, transfer and regulate music style to a certain extent. However, most of the research objects are solo music, pop music clips or symbolic music sequences, and lack of attention is paid to chorus, which is a multi-part, strong coordination and hierarchical object.

In general, the existing research has made good progress in the direction of chorus separation, automatic mixing, and style control generation, which provides a method basis and technical inspiration for this research. However, the existing methods often separate "style transfer" and "balance optimization" into different task frameworks, and lack a unified model for choral works, which can not only learn the spectral texture, rhythm density and harmonic organization of the target style, but also constrain the energy ratio between different parts, the main melody salience and the vertical hierarchical relationship during the generation process. The relevant research is summarized in Table 1.

Table 1: Related studies

Reference	Research Objective	Main Findings	Limitations
Chen et al. [4]	Improve choral music separation	Synthetic data can improve separation performance in complex choral scenarios	Focuses on separation tasks and does not involve style transfer
Jeon et al. [5]	Construct a benchmark dataset for multi-singing-voice separation	Provides a unified test benchmark and enhances result comparability	The value of the dataset is mainly reflected at the separation evaluation level
Yu et al. [6]	Achieve multi-singer voice separation based on diffusion models	Demonstrates strong disentanglement capability under zero-shot conditions	Does not address hierarchical stability after style mapping
Richard et al. [7]	Develop a differentiable unsupervised singing voice separation model	Improves end-to-end optimizability during the training process	Focuses more on source separation than on generative control
Martínez-Ramírez et al. [8]	Study deep-learning-based automatic mixing	Learns the distribution patterns of loudness and frequency bands	Lacks modeling of part organization under stylistic contexts
Koszewski et al. [9]	Build an automatic mixing system based on Wave-U-Net	Achieves relatively stable balance adjustment at the signal level	Insufficient for coordinated adaptation among multiple choral parts
Zhang et al. [12]	Explore cross-domain singing style transfer	Improves the style control capability of singing expression	Mainly targets solo singing and is not suitable for choral texture
Zhang et al. [13]	Study zero-shot singing style transfer with multi-level control	Enhances the flexibility of style control	Does not explicitly constrain balance relationships among multiple parts
Wang et al. [18]	Generate style-conditioned music based on Transformer-GAN	Verifies the feasibility of GANs in music style generation	Gives insufficient consideration to inter-part hierarchy and mixing balance

3 Methods

This paper intends to use generative adversarial networks to construct a joint optimization framework for choral work style transfer and voice part balance. Different from general music generation tasks, the object of this study is not a single melody or a solo track, but a composite musical texture composed of multiple voice parts such as soprano, alto, tenor, and bass. Therefore, the method design should not only focus on the surface acoustic features of the target style, but also simultaneously describe the entry relationship, frequency domain

distribution and energy occupancy of each voice part in the time axis. To this end, we organize multi-voice audio, corresponding music score and voice label into unified samples, and input them into the model after slicing, alignment, normalization and feature coding. In the process of style mapping, we impose content preservation constraints and voice balance constraints at the same time. The overall process is shown in Figure 1.

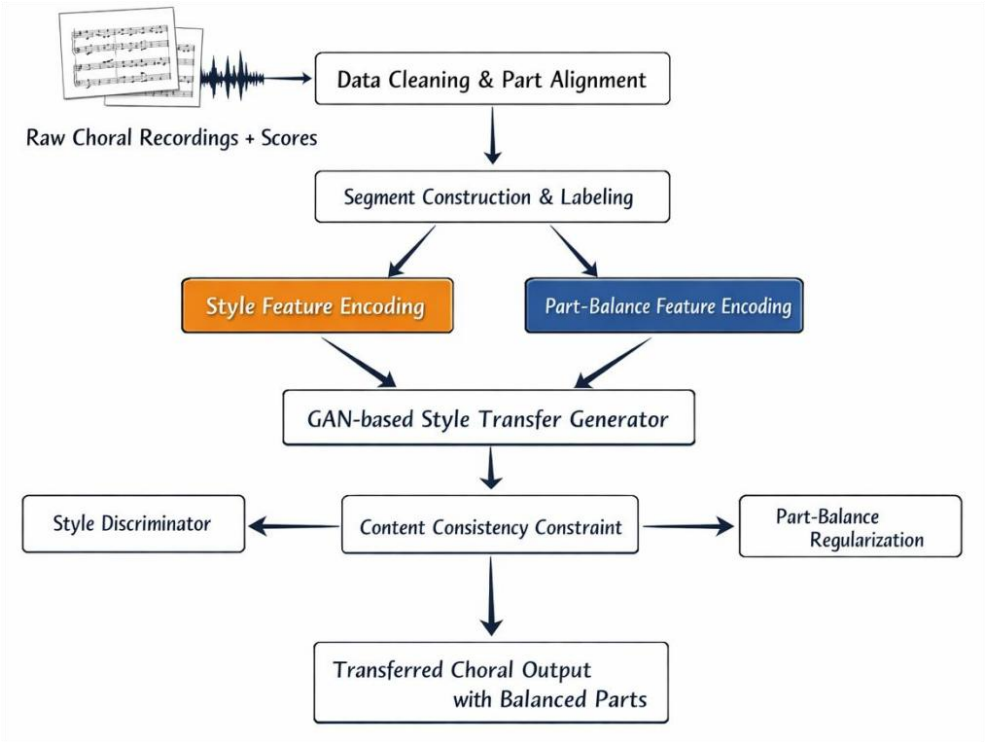

Figure 1: Overall flow chart of chorus style transfer and voice part balancing algorithm based on generative adversarial network

3.1 Data collection and sample construction of choral works

The data used in this paper consists of two parts: public multi-singing data and self-built chorus samples. In the open data part, the data resources containing multi-person singing clips, polyphonic singing clips and separation labeling information are mainly selected to enhance the basic perception ability of the model for multi-part overlapping structure. The self-built data part is collected around the mixed chorus scene, focusing on preserving the complete voice level, the difference between rehearsal version and performance version, the change of orchestration density between different styles of works, and the syntactic advancement characteristics. Considering that the task of this paper is not pure sound source separation, but joint modeling for style transfer and balance optimization, three types of information are collected synchronously in the sampling stage, such that each sample has an audible, calculable and alignable structural basis.

The data collection objects mainly cover four kinds of works: lyric chorus, sexual chorus, religious style chorus and modern adapted chorus. Standard audio files, split-part rehearsal audio, and musical scores in MusicXML or MIDI format are retained for each work, where the audio sampling rate is uniformly converted to 44.1 kHz and the quantization accuracy is set to 16 bit. In order to reduce the loudness offset and noise disturbance caused by different recording environments, the original recordings are processed by silent segment clipping, peak detection, noise threshold filtering and loudness standardization before storage. For the samples with speed fluctuation or free handling by field conductor, this paper adopts an audio

score alignment method based on dynamic time warping to map the beat point positions to a unified time grid, so as to avoid misalignment between subsequent style feature extraction and voice part balance calculation.

In the process of sample construction, the whole work is not directly used as the training unit, but the sliding slice is performed according to the phrase boundary and the harmony turning point. Let the original chorus audio be x(t) with slice length L and step size H, then the ith segment can be expressed as:

$$x_i(t) = x(t), \quad t \in [\tau_i, \tau_i + L] \tag{1}$$

The corresponding multi-part score fragment is denoted as s_i, the style label as y_i, and the voice part configuration label as p_i. Thus, the set of base samples constructed in this paper can be expressed as:

$$\mathcal{D} = \{(x_i, s_i, y_i, p_i)\}_{i=1}^{N} \tag{2}$$

where N is the total number of samples, y_i is used to indicate the style domain to which it belongs, and p_i is used to describe the activation state of each voice part of SATB in the current clip, the position of the dominant voice part and the hierarchical relationship. This definition enables the model to read both acoustic information and symbolic structure information at the same time instead of facing a single audio input when learning style mapping.

To further adapt generative adversarial training, we transcribe each segment into a unified multi-modal sample tensor. The log Mel spectrum and constant Q spectrum were extracted from the audio side, denoted as A_i. The score side encodes the pitch sequence, duration sequence, chord direction and voice part entry matrix, denoted as S_i. The integrated representation is written as:

$$Z_i = [A_i, S_i, p_i] \tag{3}$$

where p_i is the balanced prior vector of voice parts, which contains statistics such as the proportion of short-term energy of each voice part, the center of frequency band and the melody saliency. The purpose of this representation is not to simply increase the input dimension, but to provide a bridge between the "sound appearance" and the "structural skeleton" for subsequent models. If we only rely on spectrum learning, the model is easy to misjudge the local timbre change as the style core. If only based on the notation of music score, it is difficult to reflect the blending degree, thickness and light and shade changes in real singing. Multimodal sample construction is exactly able to bridge the gap between these two types of information.

In terms of sample division, this paper divides the data into training set, validation set and test set according to the principle of work-level independence, and the ratio is set to 8:1:1 to avoid information leakage caused by different segments of the same work falling into different subsets at the same time. In the case of unbalanced distribution of style categories, a resampling strategy based on class weight is used to expand the low-frequency style clips, and the original proportion distribution of voice parts is retained to ensure that the model will not mislearn a fixed voice morphology as a style feature due to sample bias. After the above processing, the final sample library not only has the differences required for cross-style learning, but also retains the internal constraints of choral works in vertical harmony, horizontal melody and voice part balance, which lays a data foundation for the subsequent training of style transfer and optimization algorithms.

3.2 Multi-voice audio and music data preprocessing

Preprocessing is a fundamental step in the modeling of choral style transfer and voice part balance. Its role is not only to organize the original data into a unified format, but also to reduce the interference caused by the difference of recording environment, singing strength, spectral writing style and time scale on model learning. For the generative adversarial network, if there are problems such as inconsistent sampling rate, offset voice entry position and non-uniform music coding granularity at the input at the same time, the model is easy to misjudge these unrelated disturbances as style differences, which will affect the stability of the transfer results. Based on this, we perform unified resampling, loudness normalization, timing alignment and symbol standardization on multi-voice audio and musical score data before model training, so that audio features and spectral surface structure can be jointly represented on the same time axis.

Audio resampling is unified with segmentation

Choral recordings from different sources differ significantly in sampling rate, bit depth and duration, and direct input to the model will increase the uncertainty of time-frequency representation. In this paper, all audio is uniformly converted to 44.1 kHz, 16 bit mono or dual channel format, and fixed-length slices are performed according to phrase boundary and beat point information. The length of the unified audio is denoted as T, and the original signal is denoted as x(t), then the segmented sample can be expressed as:

$$x_i(t) = x(t + \tau_i), \quad t \in [0, T] \tag{4}$$

Here, τ_i denotes the start time of the ith segment. Such processing helps the model to compare the spectral texture and voice part distribution of different styles of works in a fixed time window, avoiding the training instability caused by the fluctuation of segment length.

Loudness normalization

Chorus recordings often present different dynamic ranges in rehearsal hall, concert hall and studio recording environments. Without loudness normalization, the model is easy to mistake the recording level for the style attribute. In this paper, the min-max normalization method is used to linearly map the short-term energy, so that different samples are in a consistent numerical interval. Let the original amplitude be a and the normalized result be a', then:

$$a' = \frac{a - \min(a)}{\max(a) - \min(a)} \tag{5}$$

Equation (5) can compress the amplitude of each segment to the interval [0,1], thereby reducing the difference of input distribution and improving the convergence stability in the process of generative adversarial training. For the segments that need to retain the dynamic level, this paper additionally retains the original dynamic statistics after normalization, as a supplementary basis for subsequent voice part balance feature extraction.

Audio - score timing alignment

The free extension, breathing pause and conducting processing in choral works will cause the actual singing time and the spectrum beat point do not coincide completely. In order to make a one-to-one correspondence between audio spectrum frames and music score events, this paper uses dynamic time warping method to establish alignment paths. Let the audio feature sequence be $A = \{a_1, \ldots, a_m\}$, the score event sequence is $S = \{s_1, \ldots, s_n\}$, then the optimal alignment path can be written as:

$$P^* = \arg\min_{P} \sum_{(i,j)\in P} d(a_i, s_j) \tag{6}$$

Here, $d(a_i, s_j)$ represents the distance between the ith audio frame and the JTH score event. This step enables the articulation entry points, chord transition points and phrase boundaries to be accurately located in a unified coordinate system, which provides a reliable basis for the subsequent joint modeling of style features and voice balance features.

Standardized coding of music scores

On the music side, MusicXML and MIDI files are uniformly converted into four basic symbols: pitch, duration, intensity and voice part number, and encoded by discrete index to eliminate redundant differences caused by different software export formats. Marker bits were set separately for rest, extended and synchro segments to avoid the model from confusing the "silent" state with the "weak part" state. The preprocessing configuration is shown in Table 2.

Table 2: Multi-part audio and score data preprocessing configuration

Processing Stage	Processing Content	Purpose
Audio Format Standardization	44.1 kHz, 16 bit, fixed-length slicing	Ensures consistent input scale and reduces computational fluctuation
Loudness Normalization	Linearly maps amplitude to ([0,1])	Reduces the interference of recording level differences with style learning
Temporal Alignment	Matches audio frames and score events based on dynamic time warping	Ensures accurate localization of part entry and harmonic changes
Score Standardization	Unifies the encoding of pitch, duration, dynamics, and part indices	Facilitates the joint input of symbolic features and audio features
Special Marker Processing	Separately encodes rests, sustained notes, and unison states	Avoids confusion among part states

3.3 Chorus style features and voice part balance features extraction

In order to support the generative adversarial network for stable style transfer of choral works, the feature extraction link should not only stay on the general audio texture description, but also need to describe the style identification and the sound part organization law in the choral context. For this task, whether the style is effectively transferred is not only reflected by the timbre surface approaching the target domain, but also depends on whether the harmony thickness, rhythm advancement, voice texture and primary and secondary levels are still clearly discernible. Therefore, this paper jointly extracts the chorus style features and voice part balance features at the input, which are used to measure the structural consistency between the original clips and the transferred clips, and provide a computable constraint basis for the subsequent generator training. The overall extraction process is shown in Figure 2.

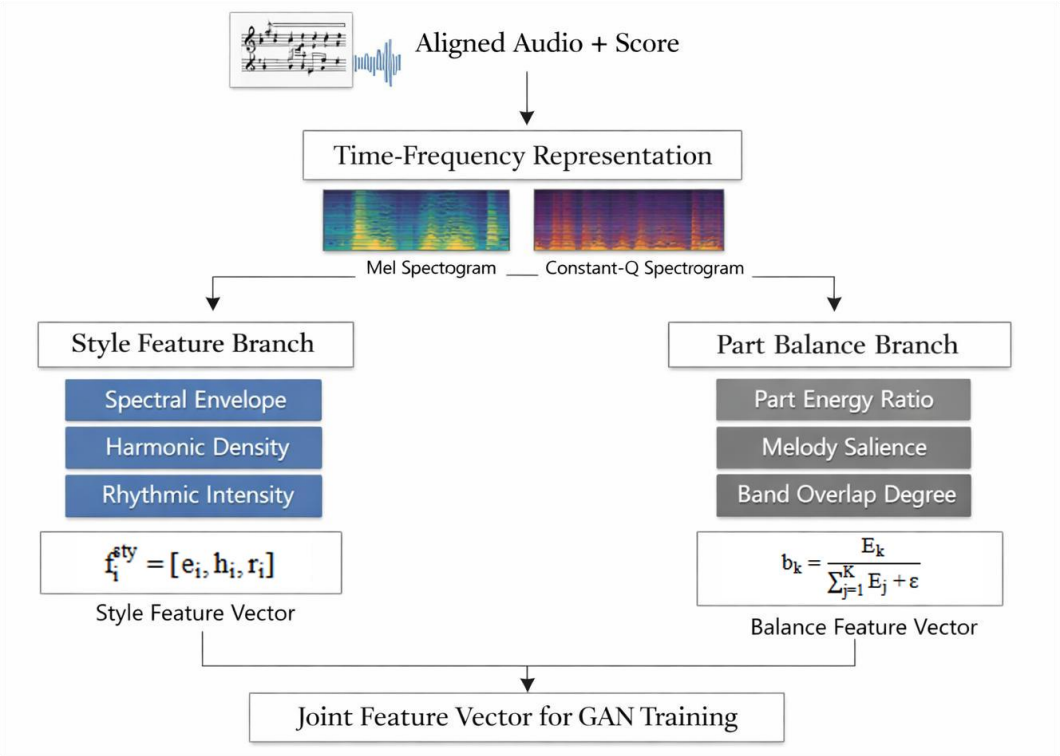

Figure 2: Flowchart of chorus style feature and voice part balance feature extraction

In terms of style feature extraction, this paper converts the aligned multi-part audio into logarithmic Mel spectrum and constant Q spectrogram, and combines the pitch direction, chord type and beat information in the music score to construct a chorus style representation vector. Let the style characteristic of the ith fragment be f_i^{sty}, then it can be written as:

$$f_i^{sty} = [e_i, h_i, r_i] \tag{7}$$

Here, e_i represents the spectral envelope feature, h_i represents the harmony density and longitudinal consonance feature, and r_i represents the rhythm intensity and syntactic fluctuation feature. This representation can completely preserve the style information of choral works at the three levels of "hearing color, harmony thickness and motion tension", avoiding the model learning shallow similarity only by local spectral texture.

In the aspect of voice part balance feature extraction, this paper focuses on the energy proportion of each voice part in a short window, the salient degree of the main melody and the degree of frequency band overlap. Let a segment contain K voice parts and the short-time energy of the KTH voice part be Ek, then its normalized energy ratio is defined as:

$$b_k = \frac{E_k}{\sum_{j=1}^{K} E_j + \varepsilon} \tag{8}$$

Here, ε is a tiny constant that prevents the denominator from being zero. Equation (8) can reflect the relative position of different voice parts in the overall choral texture, avoiding the long-term dominance of high-energy voice parts in the model update during training. At the same time, the band overlap coefficient between adjacent voice parts is further calculated to judge the phenomenon of middle and low voice parts stacking or high voice parts penetrating too strongly. If the transfer result is close to the reference sample in the target style, but the

distribution of bk is significantly shifted, it means that the model has realized the style surface level fitting, but destroyed the original chorus level.

In order to improve the comparability of features between different clips, this paper standardized the style features and the voice balance features respectively, and used the moving average method in the time dimension to suppress the occasional disturbance caused by local singing fluctuations. After this process, the model no longer receives the unsorted original spectrum, but the joint feature vector with both style discrimination ability and structure interpretation ability. This feature extraction method enables the generative adversarial network to more accurately identify "which information belongs to the style and which relationships belong to the balance" in the process of style transfer, thereby improving the integrity and audibility of the transfer results in the chorus context.

3.4 Chorus Style Transfer and Voice Part Balance Optimization Algorithm Based on Generative Adversarial Network

In order to solve the problem of style transfer of choral works under unpaired conditions, this paper constructs a bidirectional mapping model based on generative adversarial network, and introduces part balance optimization constraints into the traditional adversarial learning framework. The core idea of this method is that the model should not only learn the statistical mapping relationship between the source style domain and the target style domain, but also maintain the melody skeleton, harmony organization and hierarchical distribution of multi-part works during the transfer process, so as to avoid the phenomenon that the main voice part is masked, the low voice part is weakly supported, and the intermediate texture is overcrowded. Based on this consideration, this paper regards choral style transfer as a joint optimization process driven by "style transformation -- content preservation -- balance correction", and its overall structure is shown in Figure 3.

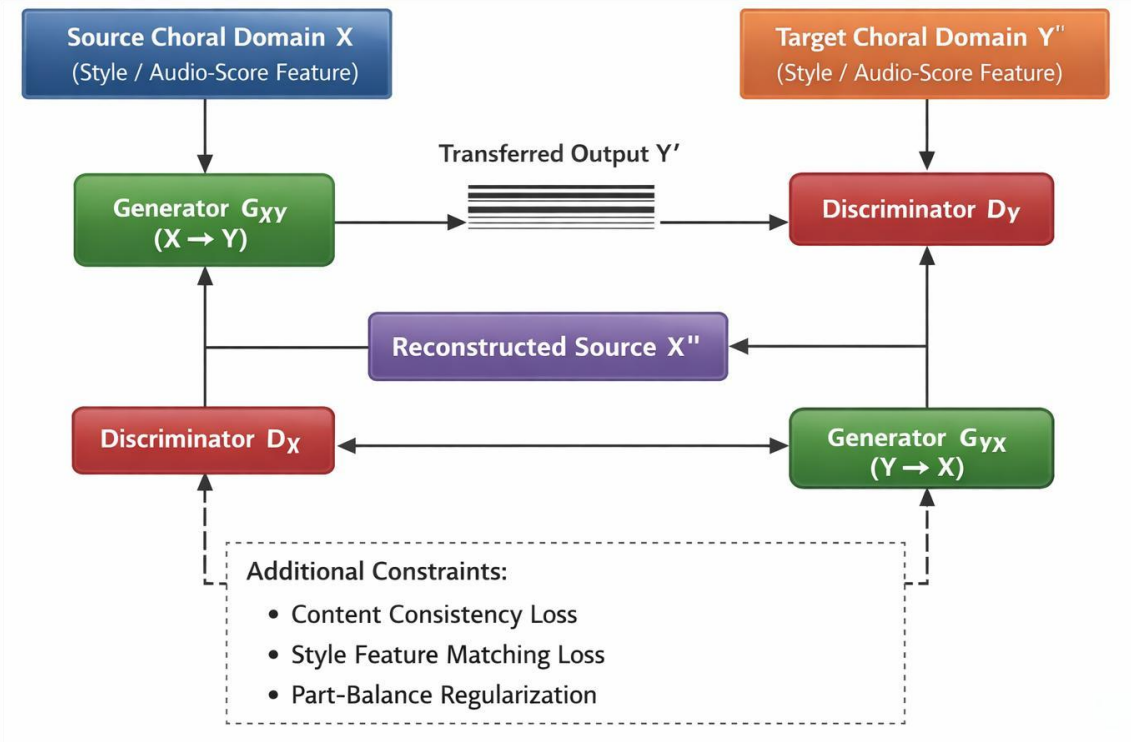

Figure 3: Optimization structure of chorus style transfer and voice part balance based on generative adversarial network

The model shown in Figure 3 contains two generators $G_{xy}: X \to Y$ and $G_{yx}: Y \to X$, and two discriminators D_y and D_x. Among them, G_{xy} is responsible for mapping the source domain choral clips to the target style domain, and Gyx is responsible for reverse mapping the target domain samples back to the source domain. Two discriminators are used to determine whether the generated results conform to the distribution characteristics of real chorus samples in their respective domains. Similar to the standard CycleGAN, this architecture is able to learn the bidirectional transformation relationship in the absence of strictly paired samples. However, this paper does not stop at the task of "inter-domain translation" in the general sense, but further emlies the voice balance features into the output of the generator and the loss function, so that the transfer results are close to the style of the target domain. It still maintains the audible ability of the choral organization. The model training process can be formulated as the following minimax problem:

$$\min_{G_{xy},G_{yx}} \max_{D_x,D_y} \mathcal{L}_{total}(G_{xy}, G_{yx}, D_x, D_y) \tag{9}$$

In Equation (9), the generator parameters are updated by minimizing the total loss, and the discriminator parameters are enhanced by maximizing the adversarial term to distinguish between real samples and generated samples. In order to balance the quality of style transfer and structure preservation, the total objective function is written as:

$$\mathcal{L}_{total} = \mathcal{L}_{adv} + \lambda_{cyc}\mathcal{L}_{cyc} + \lambda_{con}\mathcal{L}_{con} + \lambda_{bal}\mathcal{L}_{bal} \tag{10}$$

Here $\mathcal{L}_{adv}$ is the adversarial loss, $\mathcal{L}_{cyc}$ is the cycle consistency loss, $\mathcal{L}_{con}$ is the content preservation loss, $\mathcal{L}_{bal}$ is the voice part balance regularization term. λ_{cyc}、λ_{con} and λ_{bal} are the weight coefficients of each loss term. The adversarial loss is used to push the generation results to approximate the target domain distribution. Take the mapping X→Y as an example, which can be written as:

$$\mathcal{L}_{adv}^{Y} = \mathbb{E}_{y\sim p(y)}[\log D_y(y)] + \mathbb{E}_{x\sim p(x)}[\log(1 - D_y(G_{xy}(x)))] \tag{11}$$

The adversarial term of the reverse map $Y \to X$ can be defined similarly. This term ensures that the generated chorus segments are close to the target style domain in terms of spectral texture, harmonic thickness and rhythm advance. When only relying on adversarial loss, although the model may obtain the result with the target style representation, it may still destroy the original melody contour and clause structure of the source work, so it is necessary to introduce the cycle consistency constraint. Its expression is:

$$\mathcal{L}_{cyc} = \mathbb{E}_{x\sim p(x)}\left[\left\|G_{yx}(G_{xy}(x)) - x\right\|_1\right] + \mathbb{E}_{y\sim p(y)}\left[\left\|G_{xy}(G_{yx}(y)) - y\right\|_1\right] \tag{12}$$

The function of Equation (12) is to limit the sample from deviating too far from the original content after forward transfer and reverse reconstruction, so as to suppress mode collapse and irrelevant deformation. This constraint is particularly important for choral tasks, as once the reconstruction error is too large, it often means that the melodic main line, harmonic support, or part entry relationships have been severely distorted.

Considering that choral works have stronger hierarchical dependence than ordinary audio, this paper further adds a content preservation loss to impose supplementary constraints on the score skeleton and multi-voice structure. Let $\Phi(\cdot)$ denote the content representation extracted by the joint audio-score encoder, then:

$$\mathcal{L}_{\mathrm{con}} = \mathbb{E}_{\mathrm{x}\sim \mathrm{p(x)}}\left[\left\|\Phi(\mathrm{x}) - \Phi(\mathrm{G}_{\mathrm{xy}}(\mathrm{x}))\right\|_1\right] \tag{13}$$

This loss does not require the generated result to be acoustically identical to the original input on the surface, but to preserve recognizable correspondences at the level of theme direction, harmony nodes, and phrase boundaries, so that style transfer is based on "content undistorted".

In order to solve the problem of voice imbalance after chorus transfer, this paper defines the voice balance regularization term. Let the normalized energy proportion of the KTH voice part in the transfer result be $\hat{\mathrm{b}}_{\mathrm{k}}$, and the target balance reference value be $\mathrm{b}_{\mathrm{k}}^{*}$, then the balance loss is written as:

$$\mathcal{L}_{\mathrm{bal}} = \sum_{\mathrm{k}=1}^{\mathrm{K}} \left|\hat{\mathrm{b}}_{\mathrm{k}} - \mathrm{b}_{\mathrm{k}}^{*}\right| \tag{14}$$

Where K is the number of voice parts. $\mathrm{b}_{\mathrm{k}}^{*}$ is not a fixed constant, but is adaptively estimated according to the statistical distribution of similar segments in the target style domain samples. The advantage of this processing is that the model will not mechanically press all the output to a single energy template, but automatically adjust the relative weights between female high, female low, male high and male low according to the chorus habits of different styles, so that the results are more in line with the real music context.

Based on the above design, the proposed algorithm does not simply attach the style label to the generator input, but jointly shapes the chorus transfer result in four steps of generation, discrimination, reconstruction and correction. Its training process algorithm is given as follows.

Step 1: Initialize the generator G_{xy},G_{yx} with the discriminator $\mathrm{D}_{\mathrm{x}}, \mathrm{D}_{\mathrm{y}}$ parameters.

Step 2: Sample a batch of chorus clips from the source style domain X and the target style domain Y respectively.

Step 3: Extract audio spectrum, music coding and voice part balance features.

Step 4: Generate target style transfer results by G_{xy} and reverse transfer results by G_{yx}.

Step 5: Compute the adversarial loss $\mathcal{L}_{\mathrm{adv}}$ and update the discriminator.

Step 6: Compute the cycle consistency loss $\mathcal{L}_{\mathrm{cyc}}$ with the content preservation loss $\mathcal{L}_{\mathrm{con}}$.

Step 7: Calculate the balance loss $\mathcal{L}_{\mathrm{bal}}$ according to the energy proportion of each voice part.

Step 8: Jointly optimize the total loss $\mathcal{L}_{\mathrm{total}}$ and update the generator parameters.

Step 9: Repeat the iteration until the validation set loss converges.

Through this joint optimization mechanism, the model can simultaneously complete the target style absorption and voice level integration in the process of chorus style transfer, which provides a stable algorithm foundation for style similarity evaluation, balance error analysis and listening perception comparison in subsequent experiments.

4 Results and discussion

This section focuses on the experimental verification of the constructed generative adversarial network model, focusing on its comprehensive performance in both chorus style transfer and voice part balance optimization. Different from general music generation tasks, this paper not only focuses on the proximity of the transferred segments in the target style domain, but also

focuses on whether the multi-voice structure still maintains clear hierarchical relationships and recognizable melody main lines. The experimental results will be analyzed from the perspectives of training environment, parameter configuration, comparison of objective indicators and adaptability under different voice part configurations, so as to test the effectiveness and stability of the proposed method in complex chorus scenes.

4.1 Experimental environment and Dataset Settings

The experiment uses Python 3.11 as the development environment, the deep learning framework is PyTorch 2.2, the audio processing relies on Librosa and PrettyMIDI, and the running platform is Ubuntu 22.04. The hardware configuration includes an Intel Core i7-12700 processor with 32 GB of RAM and an NVIDIA RTX 4070 GPU with 12 GB of memory. This configuration can support multi-voice spectrum feature extraction, music code alignment and iterative training of generative adversarial networks, so as to ensure that samples from different style domains can complete comparable tests under the same computing conditions.

The dataset consists of public multi-singing data and self-built mixed chorus data, which includes 312 choral works, covering four styles of lyric, religious, sexual and modern adaptation. After slicing and filtering, 6840 valid sample segments are formed, each of which contains audio segments, corresponding music scores, style labels and voice part configuration information. In order to avoid information leakage caused by different segments of the same work entering the training set and the test set at the same time, this paper divides the work according to the work level, and the ratio of training set, validation set and test set is set to 8:1:1. See Table 3 for the relevant Settings. This data organization method not only preserves the cross-style differences, but also provides a unified basis for the subsequent evaluation of voice part balance error and style transfer quality.

Table 3: Experimental environment and dataset Settings

Item	Configuration
Programming Language	Python 3.11
Deep Learning Framework	PyTorch 2.2
Audio/Score Processing Libraries	Librosa, PrettyMIDI
Operating System	Ubuntu 22.04
Processor	Intel Core i7-12700
Memory	32 GB
GPU	NVIDIA RTX 4070, 12 GB
Number of Choral Works	312
Number of Style Categories	4
Number of Valid Sample Segments	6,840
Data Split	Training: Validation: Test = 8:1:1

4.2 Model parameters and training strategy Settings

In order to ensure the generation stability and the effective play of voice part balance constraints in the process of choral style transfer, the model structure parameters and training strategies are uniformly set. The generator adopted a residual encoder-decoder structure to enhance the joint modeling ability of multi-voice spectrum texture and music structure information. The discriminator uses multi-scale convolution discrimination to improve the recognition accuracy of the model for local timbral differences and the overall hierarchical distribution. The total number of rounds in the training phase is set to 120, the batch size is set to 32, Adam is selected as the optimizer, and the initial learning rate is set to 0.0002. In order

to avoid parameter oscillation in the later stage of adversarial training, the learning rate is linearly attenuated after the 80th round. In terms of activation function, ReLU is used in the generator, and Leaky ReLU is used in the discriminator to balance the nonlinear expression ability and gradient propagation stability. The weights of voice part balance loss, cycle consistency loss and content preservation loss are also set harmoniously based on the pre-experiment, and the relevant parameters are shown in Table 4.

Table 4: Model parameters and training strategy Settings

Parameter Item	Value
Number of Training Epochs	120
Batch Size	32
Optimizer	Adam
Initial Learning Rate	0.0002
Learning Rate Schedule	Linear decay after Epoch 80
Generator Structure	Encoder–Residual Blocks–Decoder
Discriminator Structure	Multi-scale Convolutional Discriminator
Number of Residual Blocks	6
Base Number of Convolutional Channels	64
Activation Functions	ReLU, Leaky ReLU
Cycle Consistency Loss Weight	10
Content Preservation Loss Weight	5
Part Balance Loss Weight	3

4.3 Performance evaluation of style transfer and voice part balance

In the performance evaluation stage, the cycle consistency loss and voice balance error are used as the core observation indicators to investigate the ability of the model to maintain the original content structure and multi-voice hierarchical relationship during the style transfer process. For the chorus task, whether the transfer result is effective depends not only on whether the target style feature is learned, but also on whether the relative strength between female high, female low, male high and male low is still in a reasonable interval. If the loss decreases and the voice error fluctuates for a long time, it means that the model has only completed the surface style fitting and has not yet established a stable choral organization ability. Figure 4 shows the trends of the two indicators during training.

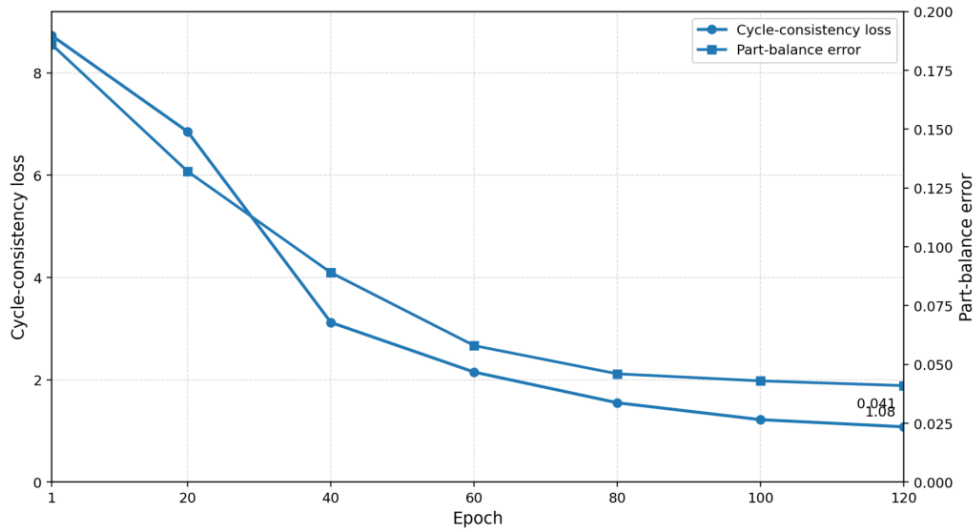

Figure 4: Cyclic consistency loss and changes in voice part balance error

Figure 4 shows that the cycle consistency loss of the model is 8.73 in the early stage of training, which has been reduced to 3.12 in the 40th round, and stabilized around 1.08 in the 120th round, indicating that the generator gradually learns to retain the melody skeleton, harmony connection and phrase boundary in the original chorus segment during the bidirectional mapping process. At the same time, the voice part balance error decreases from 0.186 to 0.041, and the fluctuation range Narrows significantly after the 70th round, indicating that the model no longer solely relies on high-energy voice parts to drive style transfer, but begins to form a more stable coordination relationship between target style absorption and hierarchical control. Both curves show a continuous downward trend, and there is no significant rebound in the later stage, which indicates that the proposed method has better convergence in the training process, and also proves that the introduced voice part balance constraint indeed improves the structural integrity and auditory balance of the transfer results.

4.4 Comparative experimental analysis

In the comparison experiment stage, the proposed algorithm is compared with basic CycleGAN, Transformer-GAN and AutoMix-Net to test the actual gain after the joint introduction of generative adversarial modeling and voice part balance constraint. The three types of comparison methods represent the bidirectional transfer model without explicit balance control, the deep generation model emphasizing conditional style generation, and the automatic mixing model focusing on energy allocation optimization. In the evaluation process, this paper mainly investigates four indicators: style similarity, main melody retention rate, voice part balance error and subjective listening score. Among them, the style similarity is calculated by the cosine similarity of the target style embedding space, the main melody retention rate is calculated by the matching proportion of the main melody track, and the subjective listening score is calculated by the 5-point MOS.

Experimental results show that the proposed method is superior to the comparison model in four indicators. The style similarity of AutoMix-Net is 0.71, the main melody retention rate is 82.4%, the voice part balance error is 0.093, and the subjective listening score is 3.78. The corresponding results of Transformer-GAN are 0.79, 86.1%, 0.068 and 4.02, respectively. The basic CycleGAN achieves 0.83, 88.7%, 0.057 and 4.15 points. In contrast, the style similarity of the proposed method is increased to 0.87, the main melody retention rate is 91.6%, the voice part balance error is reduced to 0.041, and the subjective listening score is increased to 4.36. This result shows that after adding the voice balance regularization term, the model does not weaken the style transfer ability due to the enhanced constraint, but achieves a better coordination between the target style closeness and the integrity of the chorus structure.

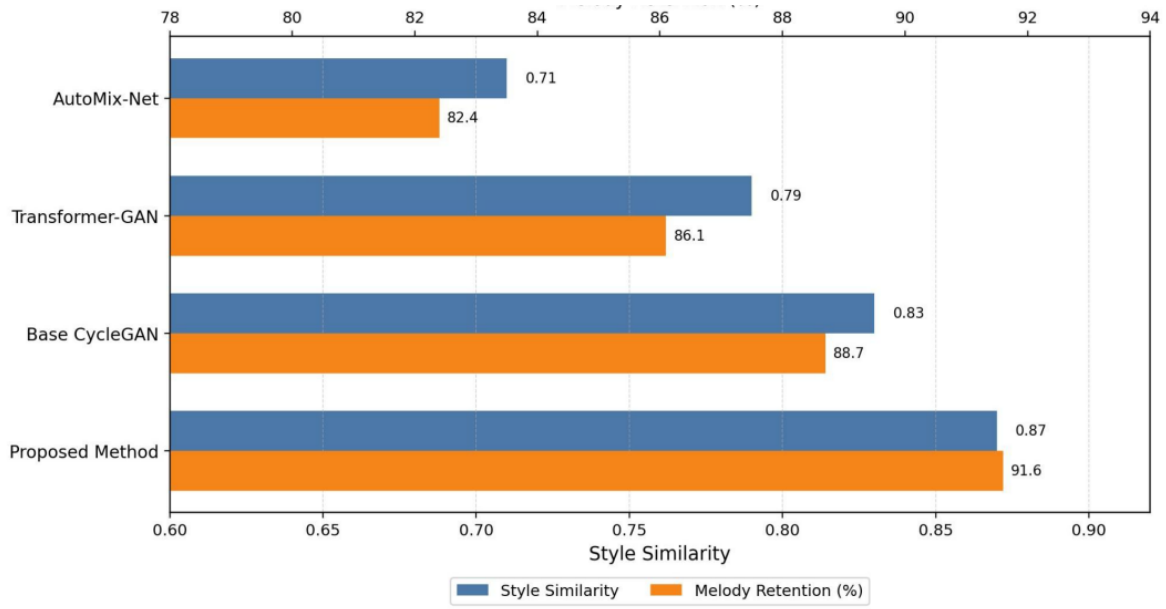

Figure 5: Comparison of style similarity and main melody retention rate for different models

It can be seen from Figure 5 that the style similarity of the proposed method is 0.04 higher than that of the basic CycleGAN and 0.08 higher than that of Transformer-GAN. In the main melody retention rate, it is 2.9 percentage points higher than the basic CycleGAN. This indicates that what the model learns is not simply spectral surface differences, but closer to the way of harmonic propulsion, texture density, and voice part relationships in the target style domain. In other words, the transfered result not only "sounds like" but is also "structurally a continuation of the original work."

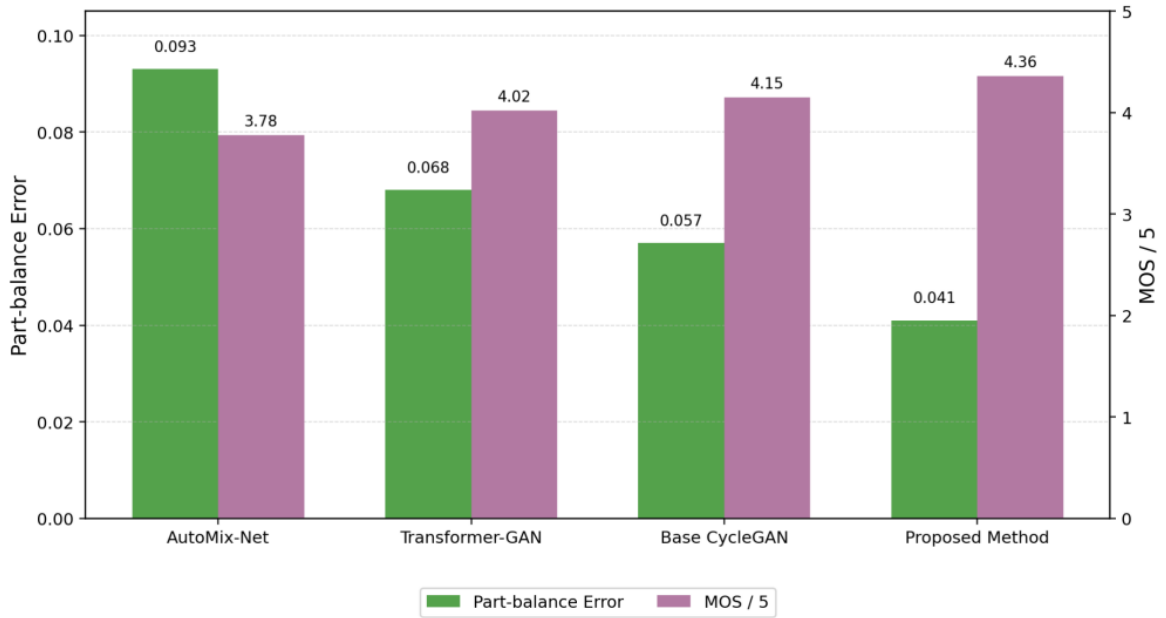

Figure 6: Comparison of voice part balance errors and subjective audibility ratings for different models

Figure 6 further shows that the voice part balance error of the proposed method is the lowest, only 0.041, which is 28.1% lower than the basic CycleGAN and 55.9% lower than

AutoMix-Net. This shows that although the traditional transfer model can complete the style domain mapping, it is still prone to the problems of bright high voice part, insufficient support of low voice part, or dull accumulation of middle voice part in the absence of explicit balance correction. By incorporating voice energy distribution, melody saliency and target style statistical features into the loss optimization at the same time, the generated results are more balanced in listening sense, so the subjective score of the proposed algorithm also reaches 4.36. Taken together, the proposed method not only improves the accuracy of choral style transfer, but also significantly improves the hierarchical stability and overall audibility of multi-part output.

4.5 Comparison of data between different methods and different voice part configurations

In order to further investigate the adaptability of the model under different chorus organization conditions, this paper divides the test data into two types according to the voice configuration. One is the standard SATB four-part mixed chorus data, where the voice boundary is clear and the relationship between the main melody and the auxiliary melody is relatively stable. The other category is the extended SSAATTBB eight-voice chorus data, which has denser harmony stacking and more obvious overlap in the middle and high frequency regions, which can better test the transfer and balance ability of the model in complex textures. On this basis, this paper calculates the style similarity, main melody retention rate and voice part balance error between the basic CycleGAN and the proposed method on the two types of configurations, and the results are shown in Table 5.

Table 5: Data comparison of different methods under different voice part configurations

Part Configuration	Method	Style Similarity	Main Melody Retention / %	Part Balance Error
SATB Four-part	Baseline CycleGAN	0.85	89.4	0.052
SATB Four-part	Proposed Method	0.89	92.3	0.036
SSAATTBB Eight-part	Baseline CycleGAN	0.79	84.8	0.071
SSAATTBB Eight-part	Proposed Method	0.84	88.9	0.049

Table 5 shows that both methods perform better on SATB four-part data than SSAATTBB eight-part data. The style similarity of basic CycleGAN reaches 0.85 in the four-part condition, and drops to 0.79 in the eight-part condition. The main melody retention rate also decreased from 89.4% to 84.8%. This result indicates that when the number of voice parts increases and the harmony texture becomes thicker, the model is more likely to suffer from the problem of too strong local texture learning and insufficient structure identification during the transfer process. In contrast, the proposed method maintains better stability on both types of configurations. On the SATB data, the style similarity reaches 0.89, the main melody retention rate is 92.3%, and the voice part balance error is only 0.036. On the eight-voice data, these three indicators are 0.84, 88.9%and 0.049, respectively, which is significantly smaller than the decline of the four-voice condition, although it is slightly lower than that of the basic CycleGAN. The comparison shows that the standard four-part configuration is more conducive to the model learning stable style mapping due to the relatively clear division of band labor and the regular entry relationship of voice parts. The extended eight-part configuration includes more complex overlapping of the same tone region and intervoice intersection, which puts forward higher requirements for the hierarchical control ability of the generative model.

4.6 Results Discussion

This paper aims to construct a generative adversarial network model for choral works, so that it can complete the style transfer while retaining the melody contour, harmonic support and voice level of the original work as much as possible. Combined with the above experiments, it can be seen that if only relying on general adversarial generation or automatic mixing methods, the model can approach the target style in local spectral texture, but it is difficult to stably deal with the problems of frequency band overlap, dominant and secondary voice competition, and vertical harmony stacking that are common in multi-part chorus. This is also the main reason why AutoMix-Net and basic CycleGAN show weakening of main melody, blurring of inner voice part and insufficient support of low voice part under complex texture condition. Transformer-GAN is superior to traditional mixing models in style representation, but due to the lack of explicit voice part balance control, its results still favor "style approximation" rather than "complete chorus structure". The above comparison results show that the advantage of the proposed method is not only reflected in the improvement of target style closeness, but also reflected in the enhanced stability of multi-voice structure under the joint action of cycle consistency, content preservation and voice part balance constraints. In this way, the generator will not excessively sacrifice the main melody skeleton and phrase boundary of the original work when learning the target style statistical features. While the discriminator pushes the results to approximate the target domain distribution, it is also indirectly constrained by the hierarchical relationship of multiple voice parts. In the experiments, the synchronous improvement of style similarity, main melody retention rate and voice part balance error of the proposed method just shows that this joint modeling strategy is effective. At the same time, the performance of the model in the eight-part condition is still slightly lower than that in the four-part condition, which indicates that there is still room for optimization of the existing feature representation and loss design when the chorus texture continues to thicken and the internal voice intersection is further enhanced. However, from the overall results, the proposed method has been able to balance the strength of style transfer and the stability of chorus organization, which has certain application value for intelligent chorus adaptation, digital music production and computational music analysis.

5 Conclusion

Focusing on the problems of main melody weakening, voice part shadowing and level imbalance that are easy to appear in the process of style transfer of multi-part choral works, this paper constructs a choral style transfer and voice part balance optimization method based on generative adversarial networks. Based on 312 choral works and 6840 valid sample fragments, this study completed the joint collection of audio and music, timing alignment, feature standardization, and the extraction of style features and voice part balance features. In addition, the cycle consistency, content preservation and voice part balance regularization terms are introduced into the bidirectional generative adversarial framework to enhance the model's ability to maintain the chorus structure. Experimental results show that the proposed method achieves 0.87, 91.6%, 0.041 and 4.36 in style similarity, main melody retention rate, voice part balance error and subjective listening score, respectively. The overall performance of the proposed method is better than that of basic CycleGAN, Transformer-GAN and AutoMix-Net. For SATB four-part configuration, the model performance is more stable. In the SSAATTBB eight-voice condition, although the index slightly decreases, it still maintains good transfer quality and hierarchical control ability. The above results indicate that the proposed method can effectively coordinate the relationship between the strength of style

generation and the organization of chorus parts. However, when the number of voice parts continues to increase and the overlap of frequency bands is further aggravated, the fine-grained control of the model still has room for improvement due to the current feature representation granularity and balance constraint form. Future research can combine the attention mechanism and more elaborate voice part prior modeling to further improve the generalization ability and auditory consistency in complex choral scenes.

Funding

This work was supported by The 2024 Research Initiation Funding Project for Introducing Doctors at Lanzhou University of Arts and Sciences, titled "Research on Choral Aesthetics and Aesthetic Education Strategies in Colleges and Universities in the New Era", with the project number 89090053.

References

[1] Wang W, Li J, Li Y, et al. Style-conditioned music generation with Transformer-GANs[J]. Frontiers of Information Technology & Electronic Engineering, 2024, 25(1): 106-120.

[2] Zhang J, Fazekas G, Saitis C. Composer style-specific symbolic music generation using vector quantized discrete diffusion models[C]//2024 IEEE 34th International Workshop on Machine Learning for Signal Processing (MLSP). IEEE, 2024: 1-6.

[3] Zhu J, Sakurai K, Togo R, et al. MMT-BERT: Chord-aware symbolic music generation based on multitrack music transformer and musicbert[J]. arXiv preprint arXiv:2409. 00919, 2024.

[4] Chen K, Dong H W, Luo Y, et al. Improving choral music separation through expressive synthesized data from sampled instruments[J]. arXiv preprint arXiv:2209.02871, 2022.

[5] Jeon C B, Moon H, Choi K, et al. Medleyvox: An evaluation dataset for multiple singing voices separation[C]//ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2023: 1-5.

[6] Yu C Y, Postolache E, Rodolà E, et al. Zero-shot duet singing voices separation with diffusion models[J]. arXiv preprint arXiv:2311.07345, 2023.

[7] Richard G, Chouteau P, Torres B. A fully differentiable model for unsupervised singing voice separation[C]//ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2024: 946-950.

[8] Martínez-Ramírez M A, Liao W H, Fabbro G, et al. Automatic music mixing with deep learning and out-of-domain data[J]. arXiv preprint arXiv:2208.11428, 2022.

[9] Koszewski D, Görne T, Korvel G, et al. Automatic music signal mixing system based on one-dimensional Wave-U-Net autoencoders[J]. EURASIP Journal on Audio, Speech, and Music Processing, 2023, 2023(1): 1.

[10] Wu X, Horner A. Diffusion Models for Automatic Music Mixing[C]//2024 IEEE International Conference on Big Data (BigData). IEEE, 2024: 3242-3247.

[11] Mukherjee S, Mulimani M. ComposeInStyle: Music composition with and without Style Transfer[J]. Expert Systems with Applications, 2022, 191: 116195.

[12] Zhang Y, Huang R, Li R, et al. Stylesinger: Style transfer for out-of-domain singing voice synthesis[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2024, 38(17): 19597-19605.

[13] Zhang Y, Jiang Z, Li R, et al. Tcsinger: Zero-shot singing voice synthesis with style transfer and multi-level style control[C]//Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. 2024: 1960-1975.

[14] Huang H, Man J, Li L, et al. Musical timbre style transfer with diffusion model[J]. PeerJ Computer Science, 2024, 10: e2194.

[15] Hong Z, Huang R, Cheng X, et al. Text-to-song: Towards controllable music generation incorporating vocal and accompaniment[C]//Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2024: 6248-6261.

[16] Sulun S, Davies M E P, Viana P. Symbolic music generation conditioned on continuous-valued emotions[J]. IEEE Access, 2022, 10: 44617-44626.

[17] Wang W, Li X, Jin C, et al. CPS: full-song and style-conditioned music generation with linear transformer[C]//2022 IEEE International Conference on Multimedia and Expo Workshops (ICMEW). IEEE, 2022: 1-6.

[18] Zhang X, Zhang J, Qiu Y, et al. Structure-enhanced pop music generation via harmony-aware learning[C]//Proceedings of the 30th ACM International Conference on Multimedia. 2022: 1204-1213.

[19] von Rütte D, Biggio L, Kilcher Y, et al. FIGARO: Controllable music generation using learned and expert features[C]//The Eleventh International Conference on Learning Representations. 2023.

[20] Ji S, Yang X, Luo J. A survey on deep learning for symbolic music generation: Representations, algorithms, evaluations, and challenges[J]. ACM Computing Surveys, 2023, 56(1): 1-39.