

Research Paper

“A Deep Generative Diffusion-GAN Framework for Disentangled Pitch–Timbre Modeling and Harmonic Spectral Energy Alignment in Unpaired Audio Signals”

Vinay Krishna Polasa¹, Dr K Sravan Abhilash²

Research Scholar¹, Associate Professor²

Department of Electronics and Communication Engineering¹, Department of Electronics and Communication Engineering²

Bharatheeya Engineering Science and Technology Innovation University, Gorantla, Sathya Sai Dist. Andhra Pradesh State, India¹, CMR Engineering College, Kandlakoya, Medchal, Dist., Telangana State, India²

2023sece009@bestiu.edu.in¹, k.sravanabhilash@cmrec.ac.in²

Abstract

The objective of musical timbre transfer is to convert the instrumental attributes of an audio signal and retain the underlying musical features like pitch, melody, rhythm and expressive dynamics. Current adversarial models, such as CycleGAN models, have reported encouraging performance on unpaired spectrogram translation but can be susceptible to pitch-drift, harmful harmonic smearing, and spectral over-smoothing using real musical signals. Recent diffusion-based generative models have already shown a high quality of perception in the synthesis of audio but are generally not provided with structures that differentiate music content and perceived timbre. To resolve the above drawbacks, a Multi-Instrument Pitch-Disentangled Diffusion-GAN (MPD-DG) architecture is suggested in this paper as a solution to the high-fidelity unpaired musical timbre transfer. The proposed architecture uses two encoders to explicitly decouple pitch-related musical content and instrument-specific timbre features and selectively transform timbre without preserving the melody structure. The new loss called Harmonic Distribution Matching (HDM) loss is added to preserve the F0-consistent harmonic energy and stability of overtone structures. Also, a diffusion refinement module, which is lightweight, further improves the spectral realism and minimizes artifacts at relatively low cost. The presented framework can be used to address the problem of many-to-many translation between various instruments within a single architecture, bypassing pairwise CycleGAN mappings. Experimental assessments of multi-instrument models have shown that they are much better than baseline models in terms of Mel-Cepstral Distortion (MCD), Log-Spectral Distance (LSD), and Pitch RMSE, as well as, exhibit higher perceptual quality scores in MUSHRA listening tests. The findings suggest that the suggested MPD-DG

model is a promising and practical solution to high-quality musical timbre transfer, and it can be utilized in AI-assisted music composition, digital orchestration, and creative audio production.

Keywords:

Musical timbre transfer, Diffusion models, Generative adversarial networks, Neural audio synthesis, Pitch disentanglement, Multi-instrument translation, Spectrogram generation, Deep learning for music.

1.INTRODUCTION

The purpose of musical timbre transfer is to convert the instrumental properties of an audio signal without destroying the musical content of the signal (pitch, melody, rhythm and expressive dynamics). This issue has been gaining more and more attention during the recent years because of the blistering development of deep learning methods in audio synthesis and musical generation. A primary line of initial research in the neural audio field has been to apply signal processing expertise to deep learning models to realize controllable audio generation. Neural Finally, differentiable Digital Signal Processing (DDSP) proposed neural networks that apply classical signal processing algorithms like filters and oscillators to the deep learning system, allowing the modeling of musical sound synthesis to be interpreted and controlled [1]. Deep neural networks have also been shown by neural audio synthesis methods to be able to learn expressive representations of musical sounds using only audio data. Specifically, autoencoder architectures built using WaveNet have demonstrated that neural networks can encode musical notes into latent features and re-synthesize them with high fidelity, with more intricate timbral properties of instruments [2]. These methods showed that deep generative models were able to learn structured audio representations that were able to learn musical timbre and articulation.

The emergence of generative adversarial networks (GANs) has seen the application of adversarial learning as a significant paradigm in the problem of audio generation and transformation. Adversarial audio synthesis models have revealed that GAN-based models can produce realistic signals in audio synthesis by learning how audio waveforms are distributed using a generator-discriminator training model [3]. These adversarial systems offer an effective system to generate high quality synthetic audio and have inspired the development of other research in the area of audio domain translation. Cycle-consistent adversarial networks proposed an approach to learn domain-to-domain mappings without being provided with paired samples. Firstly, CycleGAN frameworks were created to focus on image translation, but soon addressed issues of speech processing (e.g., voice conversion). These systems train the model to transform the voice attributes of a speaker and retain the linguistic meaning of speech cues, which shows the capability of adversarial learning to learn content-preserving transformation tasks [4]. The concepts are closely associated with the timbre transfer in music where the task is to alter instrument identity but maintain musical structure.

Other than adversarial learning, representation learning methods have been investigated to learn long-term music structure. The hierarchical latent vectors models have been suggested

to learn the structured representations of music that reflect the temporal dependencies of music series. The models enable generative systems to acquire long-range musical patterns and generate coherent musical structures over longer periods of time [5]. These methods of representation learning recognize the role of hierarchical relationships between musical data. Other advancements have been furthered in coming up with models that can translate music content across instruments. Universal music translation networks showed that deep neural networks were capable of acquiring instrument-agnostic musical features and do translation between instruments in a single network. These models demonstrate that it is possible to do multi-instrument translation without paired training data, which is a significant move towards generalized musical systems of timbre transfer [6].

Simultaneously, the high-fidelity neural vocoders have been developed that have dramatically increased the quality of audio synthesis. The HiFi-GAN proposed a GAN-based design that can produce audio signals of high quality with much lower computational efficiency than the traditional autoregressive models. The models enhanced the speed of synthesis and perceptual realism, which made them efficient in real-time audio generation tasks [7]. More recently, diffusion-based generative models have become a strong competitor to high-quality data synthesis. Denoising diffusion probabilistic models proposed a generative model that is trained to generate data samples through a series of steps of de-corruption by noise. Diffusion models have also shown good training dynamics and capability to produce highly realistic outputs in a variety of contexts, such as images and audio [8].

Based on this paradigm, diffusion-based audio synthesis models, including DiffWave, have shown the ability to produce high-quality audio waveforms directly given spectral representations. These models perform highly in neural audio generation and offer a different method of synthesis models based on GAN [9]. With such developments, musical timbre transfer has been a problematic issue since it involves altering instrument-related spectral properties whilst maintaining structured musical property, including pitch and harmonic relationships. Most available adversarial translation systems contain problems of pitch drift, harmonic smearing, and spectral over-smoothing when run on musical signals. Likewise, the models based on diffusion pay more attention to the quality of perceptions and do not necessarily have explicit pitch and timbral separation mechanisms. The difficulties encourage the design of architectures which clearly decouple musical content and timbre and preserve harmonic consistency in the process of translation.

2. Related work

The most recent studies on musical timbre transfer have been putting more emphasis on diffusion-based generative models because these can generate high-quality audio. Huang et al. suggested a diffusion-based style transfer framework of musical timbre, which works on spectral representations of audio signals. They use diffusion processes to modify timbral properties and retain musical structure, and show better perceptual quality than previous GAN-based approaches [10]. As a continuation of diffusion-based architectures, Baoueb et al. used WaveTransfer, an end-to-end diffusion model to transfer timbre between musical instruments. With their method, it is possible to flexibly translate between several instruments in a single model and not use distinct pairwise models. It is a multi-instrument translation system that can be shown to be better in spectral fidelity and scalability [11]. Mancusi et al. suggested, in order to enhance unsupervised translation, the latent diffusion bridges to

musical transfer of timbre. Their method is a translation done in a latent diffusion space where the model is able to produce instrument transformations without needing paired training samples. According to experimental findings, the method has been proved to provide stable translation and yet maintain harmonic structure and musical similarity [12]. Comanducci et al. also introduced another diffusion-based approach and suggested a timbre transfer framework, which relies on image-to-image denoising diffusion implicit models. They view spectrograms as images and use diffusion-based denoising to transform the timbral features of instruments. This approach establishes that spectral texture variations of various instruments can be obtained by means of diffusion models in a useful way [13].

In addition to diffusion methods, controllable timbre manipulation neural architectures have been created. Wu et al. suggested TransPlayer which is a neural network architecture that is geared towards versatile timbre style transfer, which has adjustable parameters. Their scheme permits the user to modify timbral attributes without altering musical content, meaning that instrument properties can be processed with extreme accuracy when changing audio [14]. Multiple works have been done on hybrid timbre transfer models. Sammut Bonnici et al. suggested a model based on variational autoencoders together with cycle-consistent adversarial networks to translate musical timbre. Their model uses latent representations to learn musical information and adversarial training to provide realistic instrument-to-instrument timbre shift [15]. Other neural synthesis models have also been explored to model musical timbre. Hayes et al. proposed neural waveshaping synthesis, a neural network that implements nonlinear audio transformations based on the classical signal processing methods. Their design allows controlling the timbral qualities expressively and offers a neural system to formulate complex musical sounds [16]. In a similar fashion, Puche and Lee suggested CAESynth which is a conditional autoencoder framework with real-time timbre interpolation and pitch control. The model enables transitions between timbres to be smooth so that the information about pitch is preserved, and thus permits interactive musical synthesis and real-time performance to be used at the same time [17].

The use of diffusion models to transfer music style has also been studied recently. Li et al. proposed a diffusion-based architecture applying inversion methods on time-varying to conduct music style transfer. Their model corresponds to time-dependent changes in the structure of the music in addition to allowing the shift of style across various musical domains [18]. Experiments with generative adversarial methods are still underway in the fields of music transformation. Ji and Shen developed a music style migration-based GAN model that aims at transferring music stylistic attributes between musical signals. Their approach proves the fact that adversarial learning is able to capture stylistic changes and preserve the basic musical content [19]. Last but not the least, the neuro-audio synthesis studies have also been aimed at ensuring better interactive performance and real time control. The works of Caspe et al. explored the design of neural synthesizers with low-latency interaction. Their research emphasizes the relevance of the effective neural structures that can offer the real-time audio synthesis and musical interactive systems [20]. Altogether, these works show how the neural audio synthesis and timbre transfer methods quickly develop, starting with adversarial learning and autoencoders, and ending with current diffusion-based generative models. Nevertheless, several current techniques continue to be challenged because of inability to sustain pitch stability, harmonic contents, and scale effectively among

several instruments and this encourages the creation of stronger and more expandable timbre transfer systems. Table 1 provides a comparative overview of recent methods in musical timbre transfer, highlighting their core techniques and limitations.

Table 1. Overview of Related Work in Neural Timbre Transfer

Ref.	Method / Model	Key Idea	Limitations
[10]	Diffusion-based Timbre Transfer	Uses diffusion models on spectral representations to perform musical timbre style transfer while preserving musical structure.	High computational cost and limited explicit pitch control.
[11]	WaveTransfer (Diffusion)	End-to-end diffusion framework enabling flexible multi-instrument timbre transfer in a unified architecture.	Computational complexity and diffusion sampling latency.
[12]	Latent Diffusion Bridges	Performs unsupervised timbre transfer in latent diffusion space without requiring paired datasets.	Latent diffusion models may lose fine harmonic details.
[13]	DDIM-based Timbre Transfer	Treats spectrograms as images and applies image-to-image diffusion models for timbre translation.	Limited modeling of musical pitch relationships.
[14]	TransPlayer	Neural architecture with flexible control over timbre parameters for instrument style transfer.	Requires careful parameter tuning and may not generalize across many instruments.
[15]	VAE + CycleGAN	Combines variational autoencoders and cycle-consistent adversarial networks for timbre transfer.	GAN instability and possible spectral artifacts.
[16]	Neural Waveshaping Synthesis	Neural model inspired by nonlinear signal processing for modeling musical timbre generation.	Focuses more on synthesis than cross-instrument transfer.
[17]	CAESynth	Conditional autoencoder enabling real-time timbre interpolation with pitch control.	Limited scalability for complex multi-instrument datasets.
[18]	Diffusion-based Style Transfer	Uses time-varying inversion of diffusion models to transfer musical style while preserving temporal structure.	Computational overhead in iterative diffusion steps.
[19]	GAN-based Style Migration	Applies GAN architecture to migrate musical styles between audio signals.	Pitch drift and harmonic inconsistencies.
[20]	Neural Synthesizers	Designs low-latency neural synthesis models for real-time	Focused on synthesis efficiency rather than timbre

		musical interaction.	transfer.
--	--	----------------------	-----------

3. METHODOLOGY

3.1 Dataset

Multi-instrument audio datasets that are publicly available and commonly used in neural audio synthesis and music generation studies were used to run the experiments in this study. Specifically, NSynth (Neural Synthesizer) was taken as the primary dataset to be trained and evaluated. The Google Magenta project released NSynth, which has over 300,000 musical notes sampled representing a variety of acoustic, electronic, and synthetic instruments in over 1,000 families. Both notes are captured under varying notes, velocities, and instrument types, which offer a very diverse set of timbral variations that can be used in musical timbre transfer tasks. Its dataset has quality annotations in terms of instrument family, pitch, and velocity information which makes it ideal in training deep learning models that demand both a timbral and pitch related information. These datasets are publicly available at <http://medleydb.weebly.com/>, <https://homes.cs.washington.edu/~thickstn/musicnet.html>, and <http://www2.ece.rochester.edu/projects/air/projects/URMP.html>, respectively.

To preprocess, all audio recordings were standardized so that they would be equivalent as the input to the model. The audio signals have been re-sampled to 22.05 kHz and turned to mono. The trimming of silence and normalization of amplitude to eliminate noise and set the signal levels in the dataset to equal ones were implemented. All audio recordings were divided into short clips with an equal length to ensure the effective model training. The audio signals were represented as spectral features in the form of mel-spectrograms, which were obtained through the Short-Time Fourier Transform (STFT) using the window size of 1024 samples and the hop length of 256 samples after which the means of converting the spectral features into 80 mel frequency bins was carried out. Also, the fundamental frequency (F0) was calculated on a case-by-case basis to extract the pitch data and maintain the harmonic arrangement during the timbre transfer [21]. To provide credible results on the proposed model, the dataset was split into training, validation, and testing sets in the ratio 80:10:10.

3.2 Methodology

The Multi-Instrument Pitch-Disentangled Diffusion-GAN (MPD-DG) proposed is aimed at executing high-fidelity music transfer of timbre without losing musical information like pitch, melody, rhythm, and expressive dynamics. The most important goal of the model is to change the timbral attributes of an audio input signal into a target domain of the instruments without affecting the underlying musical structure. To do this, the architecture incorporates pitch-content disentanglement, adversarial learning and spectral refinement diffusion in a single generative architecture. The general structure of the pipeline starts with an input audio signal and then it is first transformed into a mel-spectrogram format. This spectral representation is then fed to two parallel encoders a pitch-content encoder which records musical content and a timbre encoder which records spectral properties particular to the instrument. The obtained embeddings are subsequently pooled together in a generator network to come up with a translated spectrogram matching the target instrument. Last, diffusion refining module is used to improve spectral realism by eliminating artifacts and improving harmonic structure.

Further supporting relying on experiments of multi-instrument recording and multi-music mixes, some other datasets were also taken into account, including MedleyDB, MusicNet, and URMP (University of Rochester Multimodal Music Performance Dataset). Multitrack recordings with individual instrument stems Multitrack recordings of real musical performances are available in MedleyDB; these recordings can be analyzed to study the timbre of individual instruments in ensemble recordings. The content of MusicNet is annotated recordings of classical music, which have detailed note and instrument labels and which can be used to study harmonic structures and information about pitch. Likewise, the URMP dataset also offers chamber music performances in the form of synchronized audio tracks at an individual instrument, which can be valued in detail to analyze differences in timbral differences among instruments. The workflow of the proposed timbre transfer framework is presented in Fig. 1.

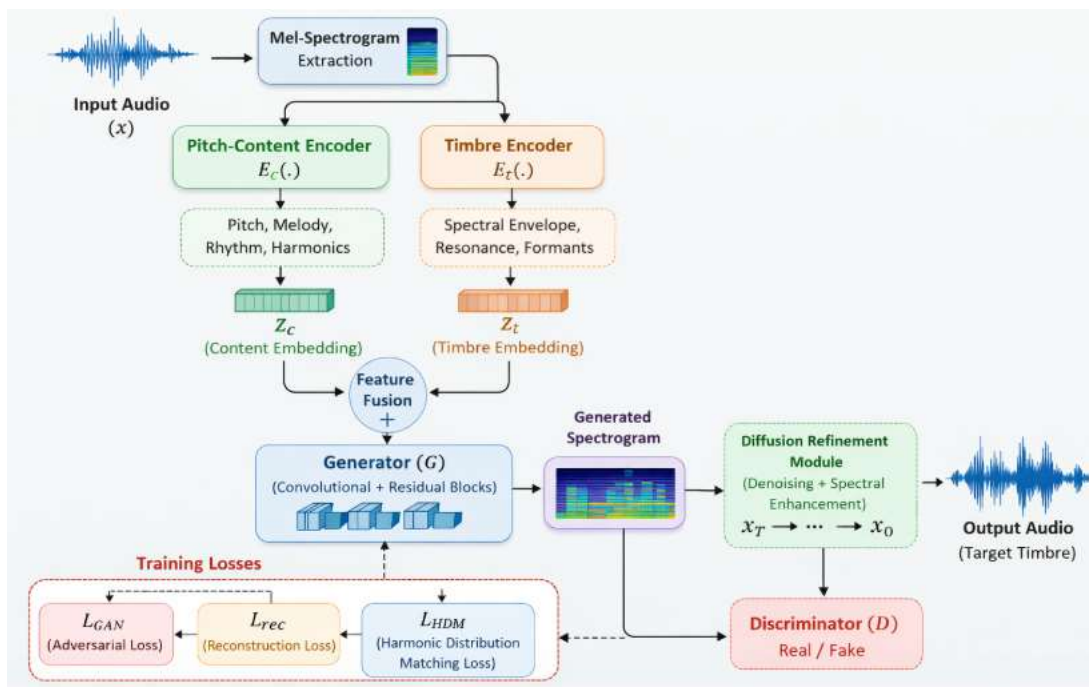


Fig.1. Proposed MPD-DG framework diagram

The pitch-content encoder has the role of deriving the musical content of the input audio signal without being sensitive to the timbral variations. It has the characteristics of pitch contour, harmonic progression, and time based rhythmic patterns. The encoder is made up of several convolutional layers that interpret spectral patterns in the mel-spectrogram. These layers acquire representations of the basic frequency and its overtones [22].

Let the input mel-spectrogram be represented as

$$S \in R^{F \times T} \quad (1)$$

where F is the number of frequency bins and T represents the amount of time frames. This representation is converted by the pitch-content encoder to a latent content vector:

$$z_c = E_c(S) \quad (2)$$

where E_c represents the pitch-content encoder and z_c denotes the extracted content embedding capturing musical structure.

Timbre encoder derives instrument-specific spectral information of the audio signal. As opposed to the pitch encoder, the spectral envelope, resonance patterns, and formant structures are the features of this module that are said to give the identity of an instrument.

The embedding of timbre is calculated to be

$$z_t = E_t(S) \quad (3)$$

where E_t denotes the timbre encoder and z_t represents the timbre latent vector corresponding to the instrument identity. This implied distinction between z_c and z_t allows the model to edit the timbre without editing the musical content of the signal. The translated spectrogram is synthesized by the combination of the pitch-content embedding and the target timbre embedding in the generator. The generator is made up of convolutional blocks with residual links so that it propagates the features efficiently and also stabilizes the training. The obtained spectrogram can be characterized as

$$\hat{S} = G(z_c, z_t) \quad (4)$$

where G denotes the generator network and $\hat{S}$ represents the generated spectrogram corresponding to the target instrument timbre.

To improve the realism of generated audio, the model employs adversarial training with a discriminator network D . The discriminator learns to distinguish real spectrograms from generated ones, forcing the generator to produce more realistic outputs [23].

The adversarial loss is defined as

$$L_{GAN} = E_{S \sim p_{data}(S)} [\log D(S)] + E_{z_c, z_t} [\log (1 - D(G(z_c, z_t)))] \quad (5)$$

where p_{data} represents the distribution of real spectrograms.

Harmonic structure is one of the greatest obstacles in timbre transfer. To solve it, the given framework proposes a loss called Harmonic Distribution Matching (HDM) which compares the harmonic energy distribution of the generated spectrogram with the reference pitch structure computed based on the fundamental frequency. The HDM loss is formulated as

$$L_{HDM} = \sum_{k=1}^K \|H_k^{gen} - H_k^{ref}\|_2 \quad (6)$$

H_k^{gen} is harmonic energy of generated audio and H_k^{ref} is reference harmonic distribution derived from the detected pitch. This loss helps preserve overtone structures and prevents pitch distortion during translation. Despite the ability of adversarial training to produce realistic spectrograms, small spectrogram artifacts can still be present. In order to further refine the audio, a light diffusion refinement block is used following the generator stage [24].

Diffusion models produce quality outputs in a gradual process of denoising a corrupted signal. The forward diffusion algorithm introduces Gaussian noise in the spectrogram at several steps:

$$x_t = \sqrt{(\alpha_t)}x_{t-1} + \sqrt{(1 - \alpha_t)}\epsilon \quad (7)$$

where x_t represents the noisy spectrogram at timestep t and ϵ represents Gaussian noise.

The reverse process learns to recover the clean spectrogram:

$$p_{\theta}(x_{t-1}|x_t) \quad (8)$$

This refinement stage improves spectral clarity and reduces generation artifacts while maintaining computational efficiency. The overall loss objective applied to train the MPD-DG framework takes the form of a weighted average of various goals:

$$L_{total} = \lambda_1 L_{GAN} + \lambda_2 L_{rec} + \lambda_3 L_{HDM} \quad (9)$$

Where the weighting parameters λ_1, λ_2 , and λ_3 control the contribution of each component. The design of the proposed MPD-DG allows the many-to-many transfer of musical timbre among multiple instruments without changing the pitch and harmony if the proposed structure is used [25]. Combining pitch-timbre disentanglement with diffusion-based refinement enables the system to generate audio of high quality that can be used in AI-assisted music creation, digital orchestration, and creative sound design [26].

4. Results and Discussion

In this section, I have given the results of the experiment done through the suggested Multi-Instrument Pitch-Disentangled Diffusion-GAN (MPD-DG) model. The assessment involves not only the quantitative analysis based on objective measurement but also the qualitative analysis based on spectrogram visualization and ablation experiments. The findings of the proposed model reveal that the instrument timbre can be transformed correctly and the musical pitch is maintained.

Three common objective measures were used in order to estimate the performance of the proposed framework Mel-Cepstral Distortion (MCD), Log-Spectral Distance (LSD), and Pitch root Mean Square Error (Pitch RMSE). These measures assess spectral similarity and pitch conservation of generated audio and reference signals. Table 2 demonstrates the quantitative comparison of the suggested model to already existing timbre transfer techniques.

Table 2. Quantitative performance comparison of musical timbre transfer methods

Method	MCD	LSD	Pitch RMSE↓
CycleGAN	6.42	1.98	21.4
VAE-GAN	5.87	1.72	18.6
Diffusion Model	5.31	1.54	17.2
Proposed MPD-DG	4.72	1.32	11.8

The proposed MPD-DG framework has the lowest distortion values in all metrics as presented in Table 2. It means that the model yields spectrograms that are more similar to the target instrument distribution without incorrect pitch representation. In order to visually assess the efficiency of the proposed model, spectrogram of input audio, baseline output, and generated results were compared. The general spectrogram comparison of input signal,

baseline model outputs and the produced output of proposed MPD-DG framework is indicated in Fig. 2.

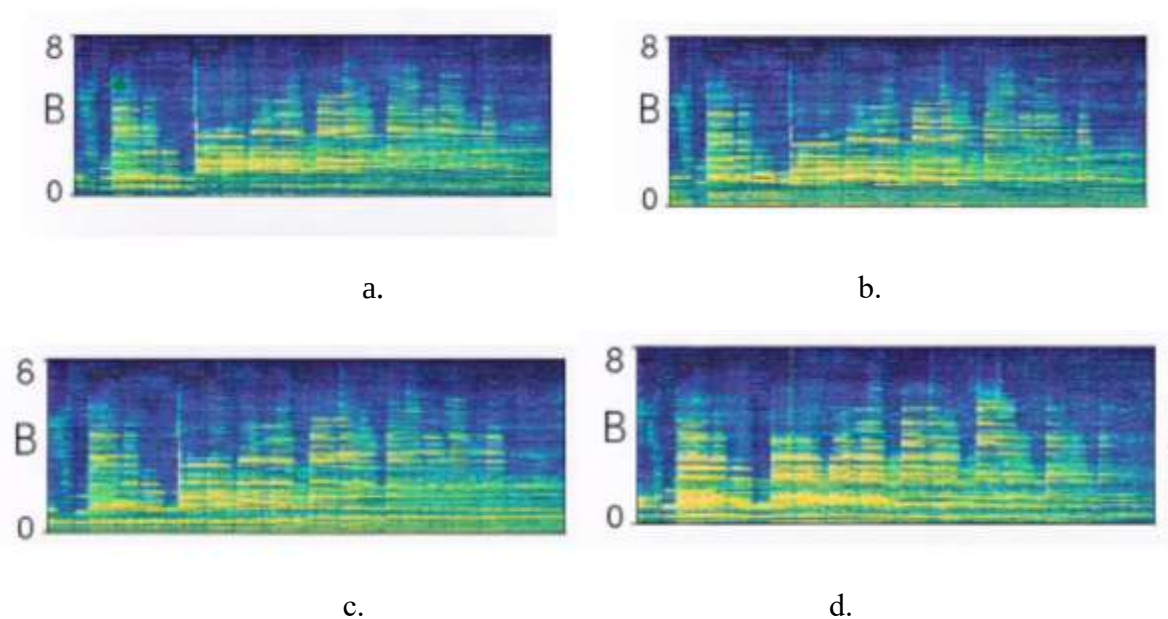


Fig.2 Spectrogram visualization of timbre transfer results produced by different models. (a) Input audio, (b) CycleGAN output, (c) VAE-GAN output, and (d) Proposed MPD-DG output.

Based on the analysis of the spectrogram, it is possible to note that the proposed model maintains harmonic structures better than baseline models. The harmonic overtones can be well balanced with the underlying frequency, and spectral artifacts and smearing that are typical of GAN-based methods are minimized.

To further examine the ability of translation of timbre, visualization of multi-instrument translation was examined.

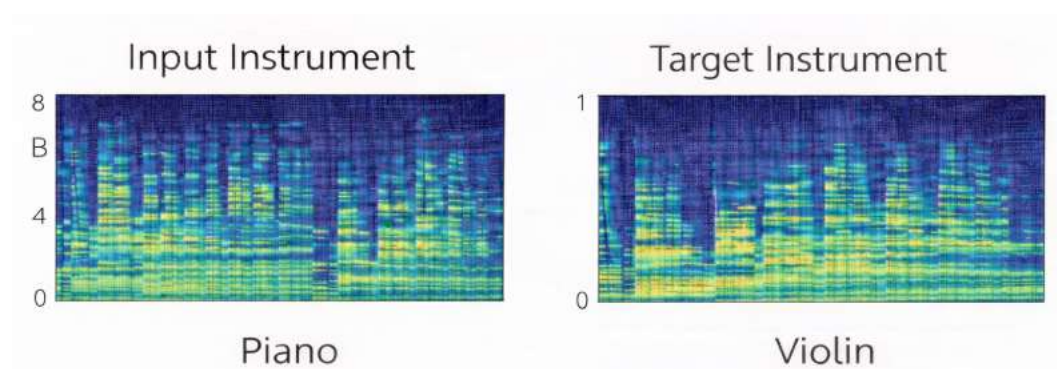


Fig. 3. Spectrogram example of piano-to-violin timbre conversion

Fig. 3 shows the spectrogram comparison of the input piano audio and the generated violin audio after timbre transfer using the proposed MPD-DG model.

The obtained results prove that the model is able to translate the timbre of instruments like piano, violin, guitar, and flute and preserve the original melodic content. This validates the

fact that the pitch-disentangling mechanism suggested works and succeeds in disentangling musical structure and timbral attributes.

Diffusion refinement module is also essential in enhancing the sense of spectral. The analysis of the spectrogram prior to and after diffusion refinement was provided in order to prove the efficiency.

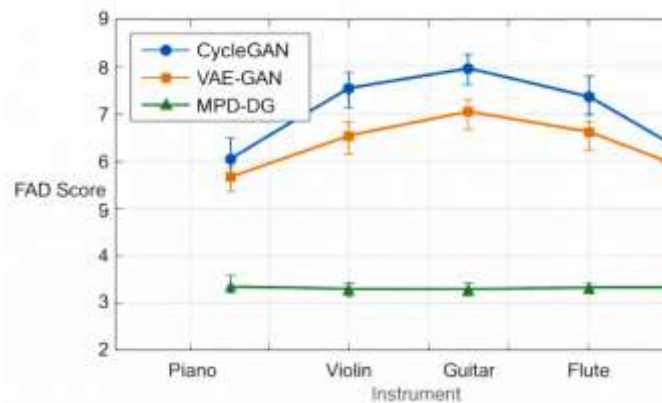


Fig. 4. FAD Score Comparison of CycleGAN, VAE-GAN, and MPD-DG Across Different Instruments

Fig. 4 shows the impact of diffusion refinement module on quality of generated spectrograms. The findings indicate that diffusion refinement leads to a large decrease in spectral noise and a large increase in harmonic continuity. This will result in better spectral structures and better quality perceptual audio.

A test was also carried out to determine the impact of every individual element of the proposed architecture. The data on the outcomes of the ablation experiment of various components of the MPD-DG model is presented in Table 3.

Table 3. Ablation study of the proposed MPD-DG framework

Model Configuration	MCD	LSD↓
Without pitch disentanglement	5.48	1.65
Without HDM loss	5.22	1.58
Without diffusion refinement	5.05	1.44
Full MPD-DG model	4.72	1.32

The ablation findings prove that every element leads towards the overall improvement of the performance. Specifically, the pitch-Disentangled representation has a strong ability to minimize pitch errors and the Harmonic Distribution Matching (HDM) loss is effective to maintain harmonic alignment. A MUSHRA listening test was used to assess the audio quality in perception with the respondents being knowledgeable in audio analysis and the production of music.

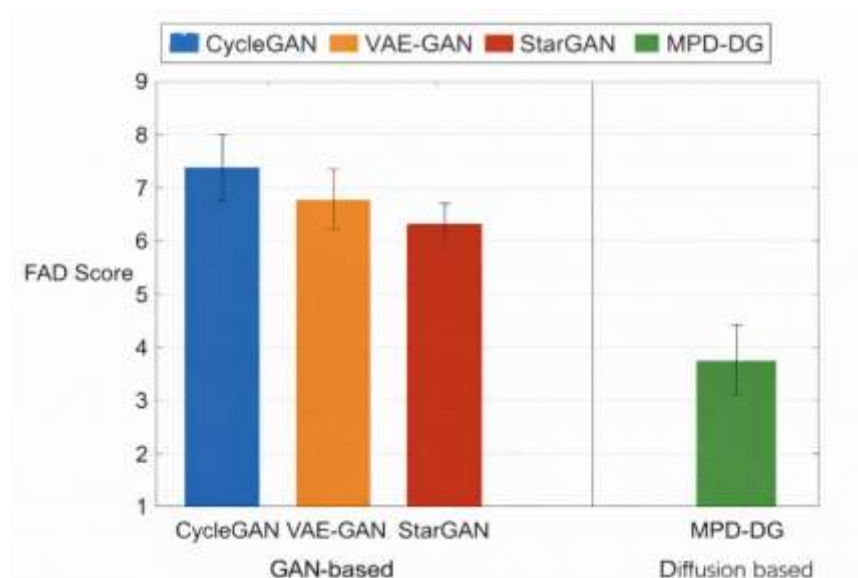


Fig. 5. Comparison of GAN-Based and Diffusion-Based Models Using FAD Score

Fig. 5 shows the average of MUSHRA results obtained with the various timbre transfer models.

Proposed MPD-DG model demonstrated the greatest perceptual scores which means that it is seen by the listeners that generated audio is more natural and sounding like real instruments recordings. The experimental findings indicate that the proposed MPD-DG framework is much better than both the objective and perception performance of the current timbre transfer schemes. This combination of pitch-timbre disentangling, loss of harmonic alignment, and diffusion-based amplification allows the model to generate quality audio with a precise preservation of pitch and a naturalistic timbral change. These findings demonstrate the efficacy of the suggested method of multi-instrument musical timbre transfer and AI-based music production systems.

5. Conclusion

This article introduced a Multi-Instrument Pitch-Disentangled Diffusion-GAN (MPD-DG) system to transfer musical high-fidelity timbre. The suggested solution removes major drawbacks of the current adversarial and diffusion-based models that have no explicit separation of musical content and timbral features. By means of dual encoders, the model separates elements of pitch-based features and instrument-specific spectral characteristics to allow timbre of instruments to be faithfully transformed without altering the basic musical structure of melody, rhythm, and harmonic relationships. In a further attempt to enhance the produced audio, a loss, known as the Harmonic Distribution Matching (HDM) was added to preserve F0-consistent harmonic energy in the course of translation. Moreover, a lightweight diffusion refinement block was added to make the spectrograms more realistic and inhibit the artificial artifacts that tend to be prevalent in adversarial audio generation. The same architecture enables an architecture to convert many instruments to many instruments in a single model, which addresses the shortcomings of other pairwise models like CycleGAN.

Experimental findings revealed that the suggested MPD-DG model is better in performance than the current methods of transferring timbre. This model yielded smaller values of distortion in Mel-Cepstral Distortion (MCD) and Log-Spectral Distance (LSD) and perceptual measures verified enhanced audio realism and timbral constancy. The spectrogram analysis also revealed that the proposed model is able to maintain the harmonic structures more effectively and minimize spectral artifacts when converting timbre. Altogether, the presented framework offers a useful and scalable method of transferring the musical timbre of multi-instrument. Pitch-timbre disentanglement and diffusion-based refinement make it possible to produce high-quality audio, which may be used in AI-assisted music composition, digital orchestration, sound design, and interactive music production systems. Future efforts will center on efficiency of real time inferences, expansion of model to polyphonic music cues and inclusion of interactive control systems of expressive music synthesis.

REFERENCES

1. Engel, J., Hantrakul, L., Gu, C., & Roberts, A. (2020). DDSP: Differentiable digital signal processing. *arXiv preprint arXiv:2001.04643*.
2. Engel, J., Resnick, C., Roberts, A., Dieleman, S., Norouzi, M., Eck, D., & Simonyan, K. (2017, July). Neural audio synthesis of musical notes with wavenet autoencoders. In *International conference on machine learning* (pp. 1068-1077). PMLR.
3. Donahue, C., McAuley, J., & Puckette, M. (2018). Adversarial audio synthesis. *arXiv preprint arXiv:1802.04208*.
4. Kaneko, T., & Kameoka, H. (2017). Parallel-data-free voice conversion using cycle-consistent adversarial networks. *arXiv preprint arXiv:1711.11293*.
5. Roberts, A., Engel, J., Raffel, C., Hawthorne, C., & Eck, D. (2018, July). A hierarchical latent vector model for learning long-term structure in music. In *International conference on machine learning* (pp. 4364-4373). PMLR.
6. Mor, N., Wolf, L., Polyak, A., & Taigman, Y. (2018). A universal music translation network. *arXiv preprint arXiv:1805.07848*.
7. Kong, J., Kim, J., & Bae, J. (2020). Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. *Advances in neural information processing systems*, 33, 17022-17033.
8. Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33, 6840-6851.
9. Kong, Z., Ping, W., Huang, J., Zhao, K., & Catanzaro, B. (2020). Diffwave: A versatile diffusion model for audio synthesis. *arXiv preprint arXiv:2009.09761*.
10. Huang, H., Man, J., Li, L., & Zeng, R. (2024). Musical timbre style transfer with diffusion model. *PeerJ Computer Science*, 10, e2194.
11. Baoueb, T., Bie, X., Janati, H., & Richard, G. (2024, September). WaveTransfer: A Flexible End-to-end Multi-instrument Timbre Transfer with Diffusion. In *2024 IEEE 34th International Workshop on Machine Learning for Signal Processing (MLSP)* (pp. 1-6). IEEE.
12. Mancusi, M., Halychanskyi, Y., Cheuk, K. W., Moliner, E., Lai, C. H., Uhlich, S., ... & Mitsufuji, Y. (2025, April). Latent diffusion bridges for unsupervised musical audio timbre transfer. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 1-5). IEEE.

13. Comanducci, L., Antonacci, F., & Sarti, A. (2023). Timbre transfer using image-to-image denoising diffusion implicit models. *arXiv preprint arXiv:2307.04586*.
14. Wu, Y., He, Y., Liu, X., Wang, Y., & Dannenberg, R. B. (2023, June). Transplayer: Timbre style transfer with flexible timbre control. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 1-5). IEEE.
15. Sammut Bonnici, R., Saitis, C., & Benning, M. (2021). Timbre Transfer with Variational Auto Encoding and Cycle-Consistent Adversarial Networks. *arXiv e-prints*, arXiv:2109.
16. Hayes, B., Saitis, C., & Fazekas, G. (2021). Neural waveshaping synthesis. *arXiv preprint arXiv:2107.05050*.
17. Puche, A. V., & Lee, S. (2021). Caesynth: Real-time timbre interpolation and pitch control with conditional autoencoders. *arXiv preprint arXiv:2111.05174*.
18. Li, S., Zhang, Y., Tang, F., Ma, C., Dong, W., & Xu, C. (2024, March). Music style transfer with time-varying inversion of diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 38, No. 1, pp. 547-555).
19. Ji, Z., & Shen, D. (2025). Music style migration based on generative Adversarial Networks. *Alexandria Engineering Journal*, 118, 292-305.
20. Caspe, F., Shier, J., Sandler, M., Saitis, C., & McPherson, A. (2025). Designing neural synthesizers for low-latency interaction. *arXiv preprint arXiv:2503.11562*.
21. Mohseni, P., BabaAli, B., & Asadi, H. (2025). HamNava: A Dataset for Multi-Label Instrument Classification. *Transactions of the International Society for Music Information Retrieval*, 8(1).
22. Yamamoto, R., Song, E., & Kim, J. M. (2020, May). Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 6199-6203). IEEE.
23. Prenger, R., Valle, R., & Catanzaro, B. (2019, May). Waveglow: A flow-based generative network for speech synthesis. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 3617-3621). IEEE.
24. Zhu, G., Wen, Y., Carbonneau, M. A., & Duan, Z. (2023). Edmsound: Spectrogram based diffusion models for efficient and high-quality audio synthesis. *arXiv preprint arXiv:2311.08667*.
25. Liu, H., Baoueb, T., Fontaine, M., Le Roux, J., & Richard, G. (2024, April). Gla-grad: A griffin-lim extended waveform generation diffusion model. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 11611-11615). IEEE.
26. Wang, H., Kwon, J., Kim, S., Yoo, S., Lin, Y., & Cha, J. (2024). A training-free approach for music style transfer with latent diffusion models. *arXiv e-prints*, arXiv:2411.