Music Genre Transfer Using Deep Generative Architectures

Shreya Sethi, Samarth Agarwal, Divij Gera, Arshia Mubias Shaik
Viterbi School of Engineering
University of Southern California
{shreyase, samartha, dgera, arshiamu}@usc.edu

Abstract

Genre transfer for music refers to the task of modifying the stylistic features of an audio sample to conform to a target genre while keeping its intrinsic musical structure intact. Earlier approaches such as CNNs and rule-based methods have been unable to generate genre-aware plausible transformations due to their limited capability in modeling long-range temporal dependencies.

In this project, we explore transformer-based models and compare them against generative models such as CycleGAN, Variational Autoencoders (VAE), StarGAN, and MusicGen. For the GTZAN dataset, we convert audio clips to Mel spectrograms and use various transformation techniques for genre transfer. We compare all models based on melody and rhythm preservation and tonal and stylistic adaptation. Among the methods explored, prompt-conditioned MusicGen demonstrated the most promising behavior in generating coherent and musically plausible outputs, indicating the potential of pretrained audio-language models for this task. Our findings offer a comprehensive empirical comparison of generative methods and demonstrate a novel use of MusicGen for real-audio genre transfer.

1 Introduction

Genre determines much of the emotion, structure, and instrumentation of music. Music genre transfer, or transforming an audio sample to have the style of a different genre, is a unique challenge for machine learning due to its need for maintaining musical structure while transforming more intangible stylistic components such as harmony, rhythm, and timbre. Existing approaches, particularly those based on convolutional neural networks or rule-based systems, fail to capture the long-range dependencies found in music.

New opportunities for addressing this task are offered by recent advances in deep generative modeling and transformer architectures. In this project, we will perform genre transfer on actual audio data from the GTZAN dataset, a common benchmark that contains music from ten genres. We implement and compare five approaches: a basic Transformer, CycleGAN, VAE, StarGAN, and MusicGen with text-based conditioning. Our evaluation focuses on the ability of each model to generate spectrograms that convincingly reflect the target genre while preserving melodic and rhythmic structure.

With qualitative analysis and audio visualization by sound, we see that while early models like the vanilla Transformer and VAE perform poorly on coherent genre transfer, StarGAN and MusicGen yield more robust solutions. Our use of MusicGen in particular showcases a novel inversion of its typical text-to-music application use case—utilizing it for prompt-based audio transformation. This project contributes to the growing literature on AI-assisted music generation and demonstrates possibilities for real-time and creative musical applications.

2 Related Work

2.1 TransGAN: Transformer-based GAN for Genre Transfer

TransGAN for music genre transfer is proposed by Kumar et al. [2022] which utilizes transformer architecture with GANs. This paper considers the task to be an image-to-image translation in Mel spectrograms to allow self-attention to preserve harmony and rhythm structure. Compared with CNN-based GANs, their work proved better genre adaptation and fewer artifacts.

2.2 MuseMorphose: Transformer-VAE based fine-grained piano music transfer

MuseMorphose, presented by Wu and Yang [2022], is a VAE model based on transformers giving style transfer in symbolic piano music on full songs and bars. The disentangled latent space learned by the model affords the fine-grained stylistic changes within constraints, such as those regarding dynamics and rhythm, with melody and musical structure preserved.

2.3 Symbolic Music Genre Transfer Using CycleGAN

Brunner et al. [2018] applied CycleGAN for symbolic music genre transfer by using piano roll representation of classical and jazzy music. The model was trained in an unpaired way and learned through cycle-consistency, while very often it created subtle transformations where changes were restricted to rhythm and note density.

2.4 Groove2Groove: One-shot symbolic style transfer

Cífka et al. [2020] described Groove2Groove, a supervised style transfer model. This model was trained using synthetic datasets and provides the conditioning of a single style reference while utilizing the rhythmic and harmonic features within the target melody. It accomplishes a rather superior content preservation and style flexibility compared with earlier GAN-based approaches.

2.5 Supervised Symbolic Style Transfer through Synthetic Data

Cífka et al. [2019] taught the encoder-decoder models to realize symbolic genre transfer on synthetic parallel datasets. By forming alignments of musical elements throughout styles, the model learned direct mappings without resorting to any human-annotated data. Very strong adaptations emerge here, with hardly any distortion of structures.

2.6 TimbreTron: Audio Style Transfer by Spectrogram Translation

Huang et al. [2023] made TimbreTron a model based on CycleGAN for audio style conversion-the process made by converting timbre through one translation of spectrograms into another. This model is, primarily, only interested in using instrument-specific tonal changes (i.e. pure pronunciation), while the instrument conversion effect (from piano to guitar) can still be achieved by the method.

2.7 MusicGen: Text-to-Music Generation with Genre Conditioning

MusicGen was initiated by Copet et al. [2023] to build a Transformer-based model that generates music from text. It has been adapted for genre transfer by conditioning generation on an input audio file and a genre-specific textual prompt. It is a first-of-its-kind application in prompt-guided style transfer using foundation models.

3 Methodology

In this project, we address the task of music genre transfer, where the stylistic aspects of a piece of music are modified to fit another genre without altering its underlying melodic and structural content. This issue has a number of challenges, including the need to model long-term temporal relationships and represent detailed stylistic characteristics such as rhythm, instrumentation, and harmony. To explore the optimal solutions, we train and compare five different generative models:

a Basic Transformer, CycleGAN, Variational Autoencoder (VAE), StarGAN, and MusicGen with prompt-based genre conditioning. The models all accept input audio as Mel spectrograms, which are extracted from the GTZAN dataset and serve as the standardized form for training and testing.

3.1 Dataset and Preprocessing

We make use of the GTZAN dataset, a popular standard for applications involving genre classification. 1,000 30-second audio songs from ten different genres—blues, classical, country, disco, hip-hop, jazz, metal, pop, reggae, and rock—make up this collection. The dataset is balanced and appropriate for genre transfer tests because each genre has 100 audio clips.

Mel Spectrogram Conversion

- All .wav audio files are converted to Mel spectrograms using the librosa library.
- Using 128 Mel bins and time steps proportional to the clip duration (about 1300 frames for a 30-second clip), each spectrogram shows the frequency content with time.
- To guarantee uniform input dimensions across models, spectograms are normalized and either trimmed or zero-padded.

These spectrograms, which act as 2D images that undergo change, are the fundamental data representation for every model in our pipeline.

3.2 Model Architectures

3.2.1 Basic Transformer

First, we implemented a standard Transformer encoder-decoder architecture to treat genre transfer as a sequence-to-sequence problem. The input is a Mel spectrogram from the source genre reshaped into a sequence of time-step vectors. The encoder transforms this sequence with multi-head self-attention layers added with positional encodings to maintain temporal properties. The decoder generates the target spectrogram autoregressively with masked self-attention and encoder-decoder attention to utilize contextual and input-dependent information. It is learned through a Mean Squared Error (MSE) between the predicted and target spectrograms. But the absence of adversarial or perceptual loss terms limits the model's capability to generate musically coherent outputs. In reality, the Transformer often failed to generate coherent audio, with the output spectrograms disintegrating into high-frequency noise and lacking identifiable structure.

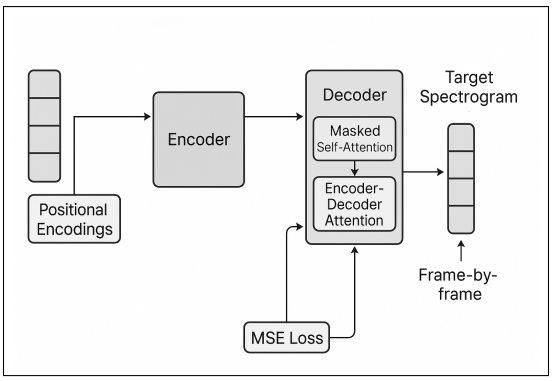

Figure 1: Basic Transformer Architecture

3.2.2 CycleGAN

CycleGAN, originally designed for image-to-image translation, is adapted here for musical genre transfer using audio spectrograms. It employs two generators (GAB and GBA) that convert spectrograms between genres A and B, built using a U-Net-style architecture with skip connections to preserve both global and fine-grained audio details. Two discriminators (DA and DB), based on the PatchGAN architecture, assess the realism of generated spectrograms at a local level. The

training process uses adversarial loss, cycle-consistency loss, and identity loss to ensure high-quality, genre-consistent audio translations.

Loss Functions Used :

- **Adversarial Loss:** Encourages generated spectrograms to resemble real ones from the target genre.
- **Cycle Consistency Loss:** Enforces $G_{BA}(G_{AB}(x)) \approx x$, preserving input content.
- **Identity Loss:** Ensures the generator behaves as an identity function when the input is already in the target domain.

Style transfer between two genres is possible with this architecture without the need for paired training data.

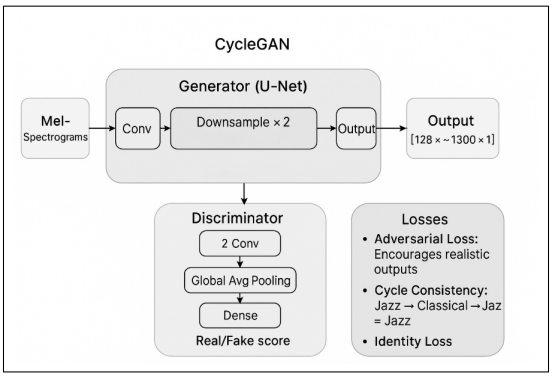

Figure 2: Cycle GAN Architecture

3.2.3 Variational Auto Encoder (VAE)

VAE is a probabilistic generative model that takes spectrograms to a latent distribution and samples from there to reconstruct new spectrograms of a target genre.

Encoder: Multiple convolutional layers down-sample the spectrogram to a latent vector. The encoder has two heads in parallel, which output the mean (μ) and log variance ($\log \sigma^2$) for the Gaussian latent distribution.

Sampling: The latent vector is sampled via the reparameterization trick:

$$z = \mu + \sigma \cdot \epsilon \quad \text{where} \quad \epsilon \sim \mathcal{N}(0, 1)$$

Decoder: The spectrogram is reconstructed from the latent vector using transposed convolutional layers.

Loss Function:

$$\mathcal{L}_{\text{VAE}} = \text{BCE}(x, \hat{x}) + \beta \cdot \text{KL}\left(q(z \mid x) \,\|\, p(z)\right)$$

where $\text{BCE}(x, \hat{x})$ is the binary cross-entropy reconstruction loss between the input spectrogram x and the reconstructed spectrogram $\hat{x}$, and $\text{KL}(q(z \mid x) \,\|\, p(z))$ is the Kullback–Leibler divergence that regularizes the encoder's posterior distribution $q(z \mid x)$ to match the prior distribution $p(z)$, typically a standard normal $\mathcal{N}(0, I)$. The hyperparameter β controls the weight of the KL term, balancing reconstruction accuracy and latent space smoothness.

3.2.4 Star GAN

StarGAN can perform multi-domain genre transfer using a single generator-discriminator pair and scale to many genres without needing to architect individual models for each genre pair. The model takes as input a spectrogram and a one-hot encoded vector of the target genre. The generator is a 9-residual block deep residual network, with the label of the genre concatenated channel-wise to condition the output. The discriminator is patch-based and produces two outputs: one for detecting

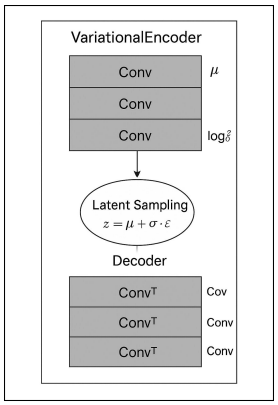

Figure 3: Variational Auto Encoder Architecture

real vs. fake samples, and the other to predict the genre class of the synthesized spectrogram. This architecture allows the model to learn genre transformation and classification simultaneously in a unified framework.

Loss Functions Used :

- **Adversarial Loss:** Encourages generated spectrograms to resemble real ones from the target genre.
- **Cycle Consistency Loss:** Ensures that transforming a spectrogram to a target genre and back yields the original input.
- **Domain Classification Loss:** Guides the generator to output spectrograms that align with the target genre label.
- **Gradient Penalty:** Enforces Lipschitz continuity in the WGAN-GP framework to stabilize training.

Overall, StarGAN offers a compact and effective architecture for many-to-many genre transfer, balancing stylistic transformation with structure preservation.

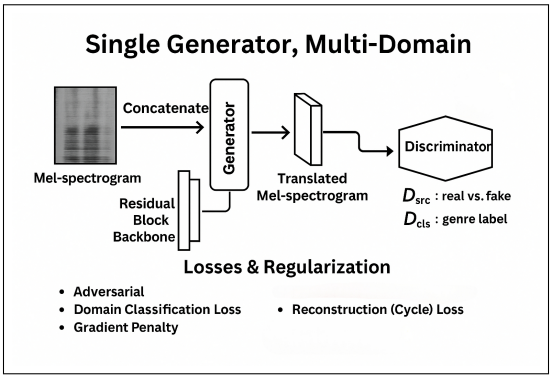

Figure 4: Star GAN Architecture

3.2.5 Music GEN

MusicGen, a pretrained transformer model developed by Meta for text-to-audio generation, is creatively repurposed in this work for prompt-guided music genre transfer. Instead of using it solely for generation, we condition the model with an input audio clip representing the source genre and a text prompt describing the desired target genre (e.g., “a classical music piece”). The input audio is first converted to a mono waveform and downsampled as part of the melody conditioning process. Simultaneously, the text prompt guides the genre style of the generated output. The result is a

30-second audio waveform that retains the melodic content of the original input while adopting stylistic elements of the prompted genre. This approach effectively harnesses the generalization capabilities of large-scale audio-language models and bypasses the need for explicitly paired training data.

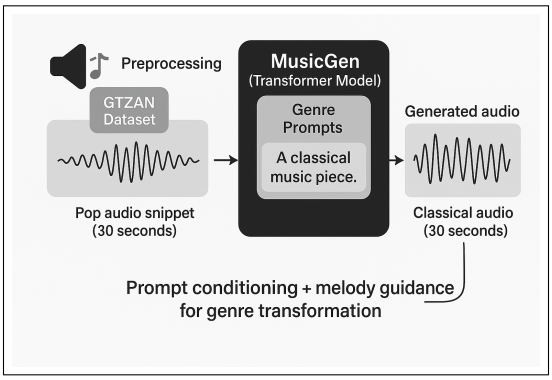

Figure 5: MusicGEN (conditioned) Architecture

4 Experiments and Results

Using the GTZAN dataset, we assess five architectures for music genre transfer: Basic Transformer, CycleGAN, Variational Autoencoder (VAE), StarGAN, and MusicGen. Based on generated spectrograms and subjective hearing, we characterize the qualitative performance of each model and pinpoint important learning and output quality constraints.

4.1 Basic Transformer

4.1.1 Observations

The Basic Transformer uses a common encoder-decoder setup trained on MSE loss with attention and positional encodings. The model was very unstable to train and suffered from bad convergence. Synthetic spectrograms were noisy and lacked any coherent frequency or time structure. The attention-based model failed to grasp the essence of genre features, and decoding was very slow due to it being autoregressive. The output samples did not display any notable genre qualities and were largely unrecognizable.

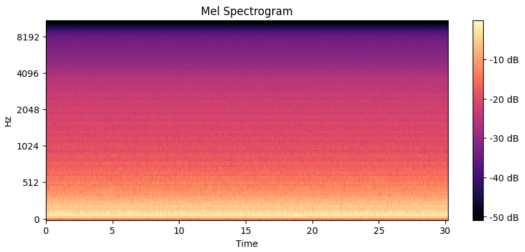

Figure 6: Input Spectrogram - Blues

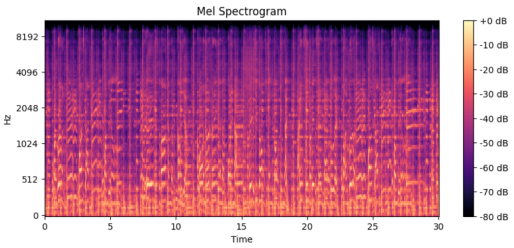

Figure 7: Output Spectrogram - Jazz

4.1.2 Limitations

- Lacks a discriminator or perceptual loss, resulting in unrealistic outputs.
- Outputs collapse to noise due to unstable training and lack of structure enforcement.
- Autoregressive decoding is slow and error-prone for long sequences.
- Mean Squared Error (MSE) loss is not ideal for audio — it fails to capture perceptual quality and musical structure.

4.2 CycleGAN

4.2.1 Observations

CycleGAN had more stable training and visually interpretable spectrograms. Cycle consistency and identity losses helped preserve temporal structure and rhythm. The transformation from source to target genres was subtle, however. The outputs were closer to the input genre than the target, and structure was preserved but stylistic features such as instrumentation or timbre were under-transformed.

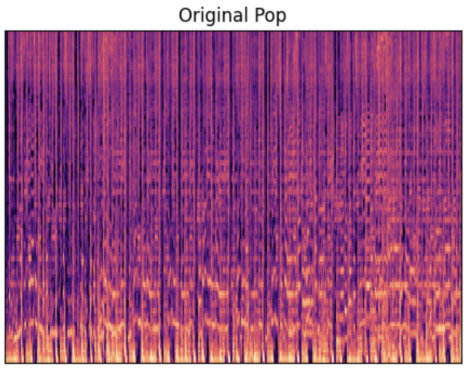

Figure 8: Input Spectogram - Pop

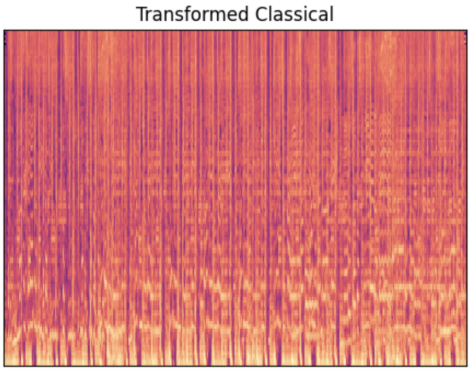

Figure 9: Output Spectogram - Classical

4.2.2 Limitations

- Genre transfer was too weak and style transfer was bad.
- It worked better with similarly paired genres and did not generalize to highly different styles.
- Suffered from lack of diversity of output and most spectrograms were just slightly different from input.
- It fixed the pairwise transfer model and needed to be retrained on new genre pairs.

4.3 Variational Auto Encoder

4.3.1 Observations

VAE performed optimally on short (2-second) segments. It was able to reconstruct simplified spectrograms with some target genre features. Latent bottleneck imposed generalization and smooth outputs. However, with whole videos (30 seconds), the model struggled to express detail. Synthetic spectrograms were too smooth and did not contain the high-contrast transients of actual music. Genre identification was bad, especially on complex audio.

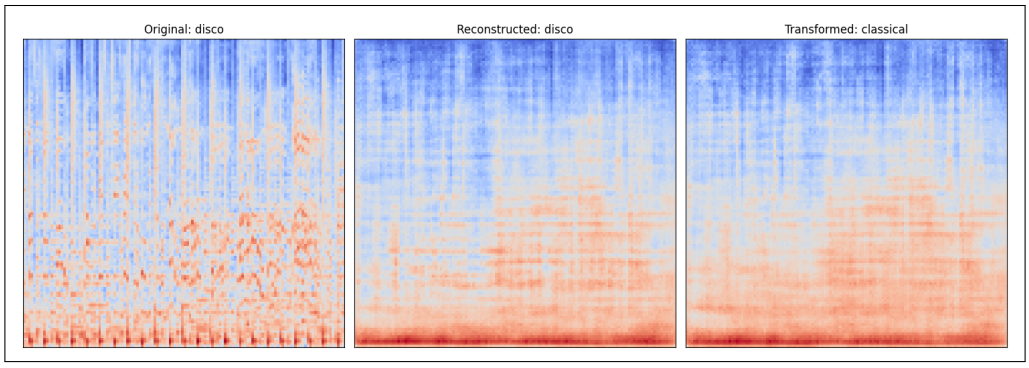

Figure 10: Disco to Classical

4.3.2 Limitations

- Poor long-term modeling: struggled with sequences longer than 2 seconds.
- Blurry outputs due to lossy compression in latent space.
- Failed to learn sharp transitions or instrument-specific features.
- One-to-one mapping only; not suitable for multi-genre transfer without retraining.

4.4 StarGAN

4.4.1 Observations

StarGAN enabled scalable multi-genre transfer using a single generator and discriminator. The model produced relatively clean and diverse spectrograms, which exposed genre-specific rhythmic and tonal characteristics. The model retained stylistic attributes such as instrumentation and beat patterns when transferring between distant genres (e.g., hip-hop to jazz). The performance was non-uniform across genre pairs; more interesting results were noticed for high-resource genres like classical and pop.

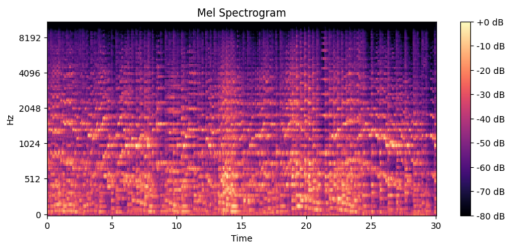

Figure 11: Input Spectrogram - Classical

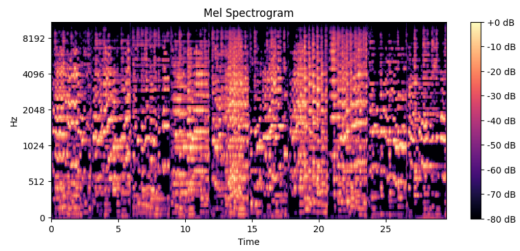

Figure 12: Output Spectrogram - Jazz

4.4.2 Limitations

- Occasional mode collapse: Certain outputs closely resembled the input genre.
- Grappled with less represented or stylistically confusing genres.
- Domain classification loss fell short of unambiguating genres in every instance.
- Outputs sometimes lacked translating consistently, generating noisy hybrid output.

4.5 MusicGEN

4.5.1 Observations

MusicGen produced the most convincing results among all models. It effectively retained the original melodic essence of the input while adopting the tonal, rhythmic, and instrumental qualities of the target genre. The process was fully automated, allowing genre transformation for any .wav file without retraining. Visualizations of the generated spectrograms confirmed distinct shifts aligned with the prompted genre, supporting the model's ability to achieve meaningful and coherent style transfer. In addition, we have created a web interface where users can convert the genres of their sounds in real time with MusicGen: they can upload an audio file and receive the transformed output with very little latency.

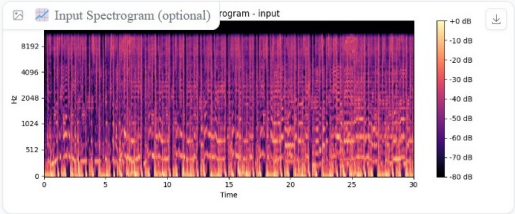

Figure 13: Input Spectrogram - Pop

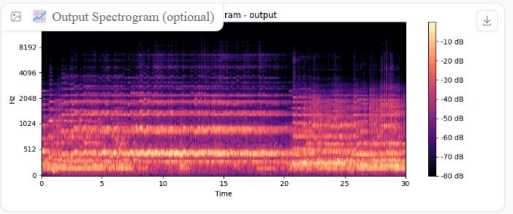

Figure 14: Output Spectrogram - Classical

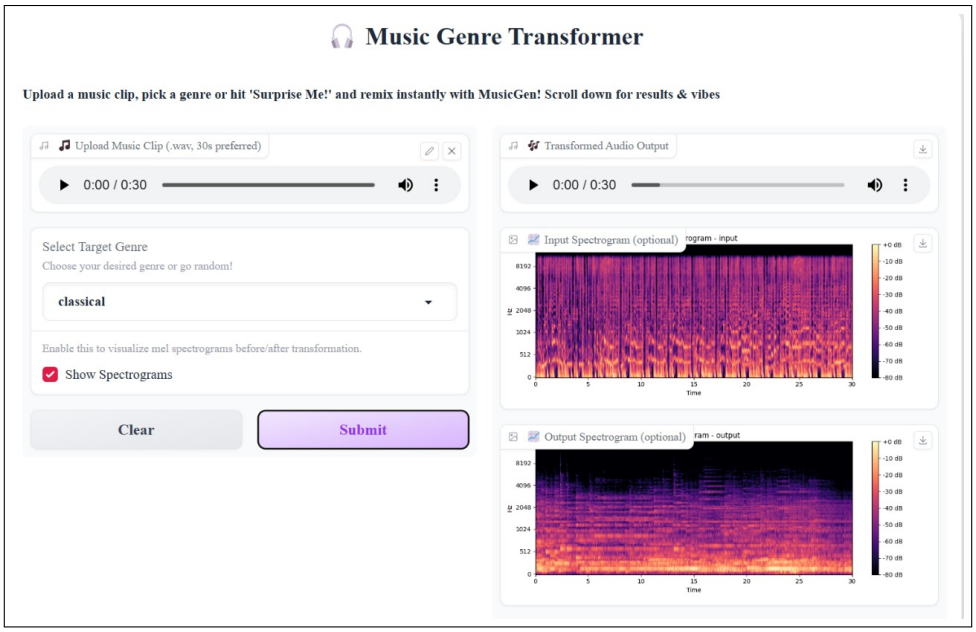

Figure 15: MusicGEN Transformation App

4.5.2 Limitations

- Lack of control for cross-breeding of genres or genre transformation strength.
- Did not perform well for certain genres like metal and reggae, especially with mixed instrumentation.
- Black-box pretraining behavior; unable to interpret or alter internal representation.

5 Discussions and Conclusions

The project delved into the rather challenging but equally stimulating space of music genre transfer, using various deep learning architectures developed from traditional transformer models to state-of-the-art conditioning-based ones. Our primary aim was the design of models that could transfer an audio sample from one musical genre to another while preserving certain salient characteristics of the music piece, such as melody, rhythm, and structure. Following a process of continuous experimentation, five different architectures were put through their paces: a simple Transformer, CycleGAN, VAE, StarGAN, and a novel adaptation of MusicGen with genre and melody conditioning. Each model had its pros and cons. Both the Transformer and the VAE architectures were theoretically sound, but they could not capture long-range dependencies that music entails and, therefore, suffered problems regarding the perceptual quality of the audio. The CycleGAN introduced adversarial training along with cyclic-consistency to preserve structure, yet the genre transfer was generally not strong. The StarGAN improved upon the multi-domain transfer itself by sending all ten genres through a single generator, though some of its outputs did show traces of the original genre. Finally, our MusicGen-based model generated impressive results, using text and melody prompt conditioning to direct the genre transformation, a reversal of the model's intended application in text-to-music generation.

The extra novelty of this work stems from giving a comparative evaluation of different architectures on the GTZAN dataset — very rarely explored outside classification — along with creatively repurposing MusicGen toward a real-audio genre transformation. The addition of a customized user interface for real-time inference lends further practical merit to the proposed solution, bringing the research closer to interactive real-world applications.

Given the challenges that remain — particularly with respect to perceptual evaluation, generalizability of the model, and handling complex genres — we have paved the way for more research. The use of transformer-based models and conditioning methods marks an important step in automating

musically aware genre transformation. The avenues that this now opens — from assisting producers in cross-genre experimentation to giving listeners personalized experiences of remixing — make this an exciting and worthwhile avenue in which to advance AI with respect to music generation.

6 Future Work

There are several promising directions for extending the current work on music genre transformation. One of the most immediate improvements involves expanding the dataset. While this project primarily relied on the GTZAN dataset, future work can benefit from using larger and more diverse audio datasets such as the Free Music Archive (FMA). These datasets contain a broader range of styles, artists, and recording conditions, which could enhance the generalization capabilities of genre transformation models and reduce overfitting on narrow genre definitions.

Another significant extension is the development of multi-genre blending capabilities. Rather than converting music from one genre to another in a binary fashion, future models could be designed to synthesize hybrid outputs that reflect characteristics of multiple genres simultaneously—such as jazz fused with electronic or classical elements. This would broaden the creative potential of generative models in music applications.

Additionally, incorporating user-controlled transformation mechanisms could offer a more interactive and customizable experience. Features like genre strength sliders or conditioning vectors would allow users to influence the degree to which the target genre affects the final output, enabling more nuanced control over the transformation process. This could be particularly useful in creative or educational music tools.

Further fine-tuning of the MusicGen model also presents a valuable direction. By training on more domain-specific datasets or extending conditioning sequences, it may be possible to improve both genre fidelity and melodic coherence, leading to more convincing and musically structured outputs.

Lastly, a valuable area of exploration lies in incorporating multimodal inputs. Beyond audio, integrating metadata such as lyrics, tempo, or even mood annotations could help inform genre transformation in a more expressive and context-aware manner. This would allow the model to generate outputs that are not only genre-appropriate but also emotionally and thematically aligned with the original content.

Together, these directions offer a rich roadmap for improving the quality, flexibility, and creativity of music genre transformation systems in future research.

References

G. Brunner, Y. Wang, R. Wattenhofer, and S. Zhao. Symbolic music genre transfer with cyclegan. *arXiv.org*, September 20 2018. URL `https://arxiv.org/abs/1809.07575`.

Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Défossez. Simple and controllable music generation. *arXiv preprint arXiv:2306.05284*, June 2023.

O. Cífka, U. Şimşekli, and G. Richard. Supervised symbolic music style translation using synthetic data. *arXiv preprint*, arXiv:1907.02265, 2019. doi: 10.5281/zenodo.3527878.

O. Cífka, U. Şimşekli, and G. Richard. Groove2groove: One-shot music style transfer with supervision from synthetic data. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:2638–2650, 2020. doi: 10.1109/TASLP.2020.3019642.

Sicong Huang, Qiyang Li, Cem Anil, Xuchan Bao, Sageev Oore, and Roger B. Grosse. Timbretron: A wavenet(cyclegan(cqt(audio))) pipeline for musical timbre transfer, 2023. URL `https://arxiv.org/abs/1811.09620`.

S. Kumar, J. Verghese, and R. Dutta. Music genre transfer using transgan. In S. Smys, V. E. Balas, and R. Palanisamy, editors, *Inventive Computation and Information Technologies*, volume 336 of *Lecture Notes in Networks and Systems*. Springer, Singapore, 2022. doi: 10.1007/978-981-16-6723-7_63.

S.-L. Wu and Y.-H. Yang. Musemorphose: Full-song and fine-grained piano music style transfer with one transformer vae. *arXiv.org*, December 19 2022. URL `https://arxiv.org/abs/2105.04090`.