BAND-TO-BAND STYLE TRANSFER VIA UNIVERSAL MUSIC TRANSLATION NETWORK: HEARING DAY6 AS SILICAGEL

Jooeun Lim
Graduate School of Culture Technology, KAIST
jooeun9199@kaist.ac.kr

Taehui Lee
Graduate School of Culture Technology, KAIST
taehui@kaist.ac.kr

ABSTRACT

We propose a band-to-band music style transfer model using a WaveNet-based universal music translation architecture. By training artist-specific decoders and a shared encoder, we explore domain transfer in long-form instrumental music. We evaluate the latent space, loss convergence, and inference behavior, and discuss limitations and future work involving diffusion-based models.

1. INTRODUCTION

Music style transfer aims to transform the stylistic characteristics of a musical piece—such as instrumentation or mixing—while preserving its underlying content. While early studies focused on symbolic representations [2], recent works like the Universal Music Translation Network (UMTN) [9] have demonstrated audio-domain transfer using shared latent spaces and artist-specific decoders.

In this project, we explore band-to-band style transfer between real instrumental recordings from six Korean bands, including DAY6 and Silicagel. Unlike voice conversion [11] or Foley generation [7], our task involves long-form audio.

We adopt a WaveNet autoencoder architecture as our baseline and adapt it to a multi-domain setting, training a shared encoder and artist-specific decoders across six bands. Our experimental analysis includes latent vector visualization, loss convergence, and inference time. Additionally, we explore fine-tuning strategies to improve stylistic fidelity and discuss the potential of diffusion-based models, such as MusicLDM [3], for future work.

Our contributions are as follows:

- We construct a multi-artist audio dataset consisting of instrument-only tracks from six Korean bands, with no vocal content.
- We adapt and fine-tune a universal music translation network to the band-to-band style transfer task.
- We analyze the model's representation and generation capabilities and highlight key limitations and future directions.

2. RELATED WORK

2.1 Style Transfer

Style transfer has been actively explored in the vision domain, where early works predominantly relied on generative adversarial networks (GANs), including StyleGAN [6], CycleGAN [14], and AdaIN [5]. More recently, text-conditioned diffusion models [12] have demonstrated superior performance in both quality and controllability, leading to a shift in dominance from GAN-based methods. Approaches such as classifier-free guidance and text-conditioned sampling [8, 12] have become standard practice in diffusion-based generation. These methods, originally popularized in vision applications, have since been successfully extended to the audio domain.

In contrast, style transfer in the audio domain has been relatively limited, with existing studies focusing mainly on short-duration signals, such as Foley sound synthesis [7] or voice conversion [11]. These tasks typically operate on brief audio segments and do not address long-form structure.

Our work extends style transfer to the musical audio domain, with a particular focus on long-form music. This involves capturing both local timbral characteristics and global structural coherence, which are critical to realistic music transformation. A related approach is the Universal Music Translation Network [9], which addressed instrument-style transfer in a domain-agnostic fashion, but did not consider genre transfer or long-form generation.

2.2 Music Genre Transfer

Existing methods for music genre transfer primarily adopt CycleGAN-based architectures [2], and most operate within the symbolic domain, such as MIDI. While symbolic representations reduce temporal complexity, they lack the rich acoustic detail of real audio. Consequently, such approaches fall short in modeling the expressive nuances of genre-specific musical textures.

In this study, we propose an end-to-end framework that directly operates on audio waveforms, enabling genre transfer at the waveform level. This allows us to fully leverage the diversity and complexity present in real-world audio data, facilitating more faithful and expressive style adaptation.

2.3 Audio Representation Network

WaveNet [13] was a pioneering autoregressive model for raw audio generation, inspiring subsequent architectures such as WaveGAN [4] and WaveGlow [10], which explored GAN-based and flow-based paradigms respectively. While these models achieved notable success, they are limited in scalability and parallelism.

Following recent advances in vision and language domains, diffusion models have become the leading approach in audio generation. Models such as AudioLDM [8] and Stable Audio [1] have shown impressive capabilities in generating high-quality audio from text prompts, paving the way for more controlled and expressive music synthesis tasks.

3. METHOD

Style transfer, originally studied in the visual domain, has recently been extended to music with the advent of deep generative models. Among these, GAN-based approaches such as CycleGAN [14] have been widely used for musical style transfer tasks due to their ability to perform unpaired domain translation.

However, applying such methods to real-world music introduces significant challenges. GAN training is highly unstable and sensitive to hyperparameters and data diversity. Furthermore, musical signals contain long-range temporal dependencies—such as rhythm patterns, phrase structure, and dynamic progression—which are difficult for GANs to model consistently, especially in an autoregressive context.

To address these limitations, we adopt a WaveNet autoencoder architecture designed for universal music translation. Rather than relying on adversarial learning, this model encodes musical content into a shared latent space via a convolutional encoder and reconstructs domain-specific output using separate WaveNet decoders. Its design supports flexible multi-domain representation while preserving fine-grained temporal features, making it well-suited to our task: translating the band sound of Day6 into the stylistic space of Silicagel.

Although our motivation for using this architecture was partly pragmatic, it also aligned with our objective. Since vocals were excluded from the input, the focus was on capturing differences in instrumentation, spatial depth, and production style between the two bands. The chosen architecture enabled us to explore these transformations effectively within a unified and domain-aware framework.

3.1 Baseline Architecture: WaveNet Autoencoder for Multi-Domain Music Translation

Our project builds upon an existing WaveNet autoencoder architecture designed for universal music translation across multiple musical domains. The model is composed of a shared convolutional encoder and domain-specific autoregressive decoders based on WaveNet, as shown in Figure 1. This setup enables the model to learn high-level representations of audio content that are independent of

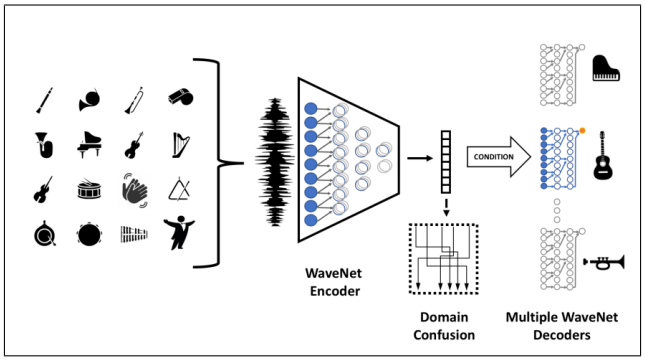


Figure 1: The architecture of baseline network.

specific stylistic domains, while still allowing for domain-specific reconstruction via individualized decoders.

The encoder is a fully convolutional network consisting of 30 residual dilated convolution layers, grouped into three blocks. Each residual layer includes ReLU activations and 1×1 convolutions, followed by an average pooling operation with a window size of 50 ms, resulting in a temporally downsampled latent vector in $\mathbb{R}^{64}$. This latent vector serves as a bottleneck representation that captures the semantic content of the input audio.

To guide the model away from simply memorizing low-level audio features, a pitch-based augmentation scheme is applied during training. Short segments (0.25–0.5 seconds) of each 1-second input are pitch-shifted randomly using Librosa, encouraging the encoder to focus on robust, domain-invariant representations.

Each decoder is a domain-specific WaveNet conditioned on the encoder output. It consists of four blocks of 10 residual layers, giving it a receptive field of 250 ms. The decoder generates mu-law quantized 8-bit audio samples in an autoregressive manner. The conditioning vector is upsampled to the original audio rate and passed through dedicated 1×1 convolution layers for each decoder layer.

To regularize the latent space, the model includes a domain confusion network trained to predict the original domain label from the encoder's output. During training, this network contributes an adversarial loss that penalizes the encoder for leaking domain-specific information, encouraging better generalization across unseen musical styles.

At inference time, the transformation from one musical domain to another is performed by encoding the input audio using the shared encoder and decoding it with the target domain's decoder, without applying any augmentation.

3.2 Dataset Construction and Domain Adaptation

To apply the baseline WaveNet autoencoder model to our band-to-band style transfer task, we constructed a custom multi-domain dataset consisting of instrumental tracks from six Korean alternative or rock artists: Day6, LUCY, Car, the garden, Jannabi, Thornapple, and Silica Gel. Only instrumental versions were used to eliminate the influence of vocal characteristics and focus purely on differences in instrumentation, mixing style, and spatial production.

Audio tracks were sourced from publicly available YouTube videos and were used for training after being randomly sliced into fixed-length segments at a consistent sampling rate. Since these tracks did not include vocals, no explicit vocal removal or source separation was necessary. Each artist was assigned to a distinct domain label, resulting in a six-domain classification setup for training and domain confusion.

The domain-agnostic encoder and the artist-specific decoders allowed us to explore mappings between distinct musical identities. In particular, the model was trained to reconstruct a given artist's instrumental sound from its latent representation, while simultaneously learning a shared feature space that abstracts away from domain-specific cues.

3.3 Fine-tuning Strategies and Model Adaptation

Following the initial training of the baseline model, we conducted a series of fine-tuning experiments to improve stylistic alignment and reconstruction quality across all six artist domains. Rather than focusing on a single pairwise transformation, we applied fine-tuning procedures individually to each domain to better capture its unique stylistic characteristics. The following strategies were explored:

1. **Decoder-Only Fine-Tuning**: The shared encoder was kept fixed while each artist-specific decoder was retrained using that artist's own data. This tested the decoder's ability to adaptively reinterpret a fixed latent space.

2. **Full Fine-Tuning (Encoder + Decoder)**: Both the encoder and each domain-specific decoder were jointly fine-tuned using the corresponding artist's data. This allowed the latent space itself to shift toward representations more appropriate for each domain.

3. **Decoder Re-Finetuning**: Building on the full fine-tuning setup, we further trained only the decoder while freezing the encoder. This assessed whether the decoder alone could refine outputs once the latent space had already been adapted.

Model performance was evaluated through qualitative listening tests, as well as loss curve monitoring and latent space visualization after encoder fine-tuning. These evaluations were used to assess perceptual coherence and representation shift, but quantitative metrics were not applied at this stage. A detailed comparison of outcomes is presented in the Results section.

4. EXPERIMENTS

4.1 Dataset

We collected instrument-only tracks from YouTube for six Korean bands: DAY6, Silicagel, Car, the Garden, Jannabi, Lucy, and Thornapple. The tracks were downloaded using `youtube-dl`. The total durations per band are as follows: DAY6 (2h 30m), Silicagel (47m), Car, the Garden (47m), Jannabi (1h 4m), Lucy (4h 52m), and Thornapple (3h 16m). Each dataset was used to train a separate domain-specific decoder in the universal translation framework. Each batch consisted of samples from a single artist, and artists were cycled in a fixed order across batches to ensure balanced training. For each batch, a random song from the selected artist was chosen, and a random segment was extracted to encourage intra-domain variation and temporal diversity.

4.2 Experimental Settings

The training configurations for each experiments were:

1. **Decoder-Only Fine-Tuning**: The decoder was fine-tuned for 10 epochs using 100 iterations, with a learning rate of 1e-4. Training was performed on a single NVIDIA RTX 3090 GPU with a batch size of 16.

2. **Full Fine-Tuning (Encoder + Decoder)**: Both encoder and decoder were trained from pretrained weights for 64 epochs (1000 iterations). The encoder learning rate was set to 1e-3 and decoder to 1e-4 to balance the update count between the shared encoder and the decoders. This setting used two NVIDIA RTX 4080 GPUs with a batch size of 8.

3. **Decoder Re-Finetuning**: Additional fine-tuning of the decoder was performed after configuration (2) for 10 epochs, using the same hyperparameter and hardware setup as in (1).

4.3 Results

4.3.1 Latent Vector Distribution.

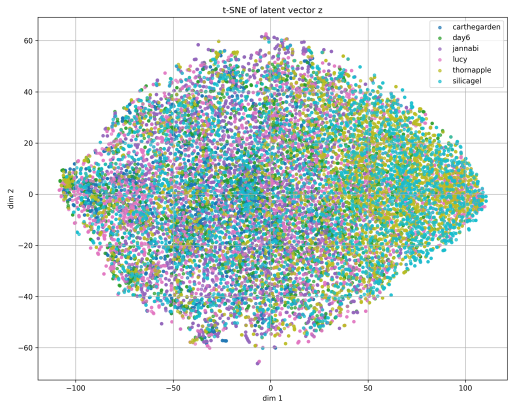
As training progressed, the latent space exhibited more dispersion, indicating that the model learned to encode varied musical characteristics. However, the clusters representing different bands were not distinctly separable in the projected space, as shown in Figure 2. Considering that the latent vectors are derived from short, time-independent audio segments, this partial separation is still encouraging. Additional temporal conditioning or contrastive learning objectives may help promote better class-wise separation in future work.

4.3.2 Loss Performance.

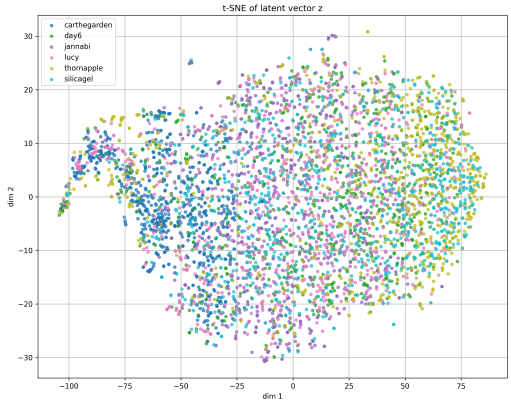
Training loss continued to decrease over time, confirming effective convergence, as shown in Figure 3. However, the model's ability to distinguish styles across bands remained relatively unchanged, suggesting that further training or architectural enhancements might be required.

4.3.3 Inference Time.

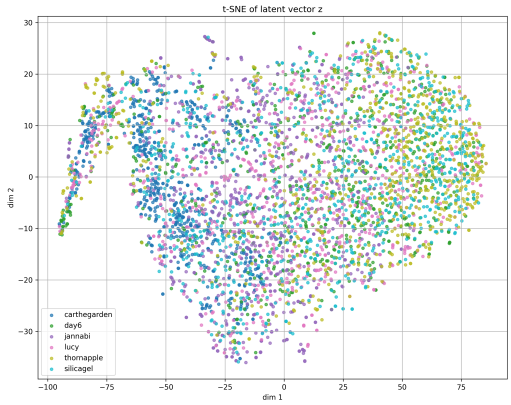
Generation of a 5-second sample took approximately 2.9 minutes, which is impractical for real-time applications. The bottleneck arises primarily from the autoregressive nature of the WaveNet decoder. Replacement with a faster, parallel generation module, such as flow- or



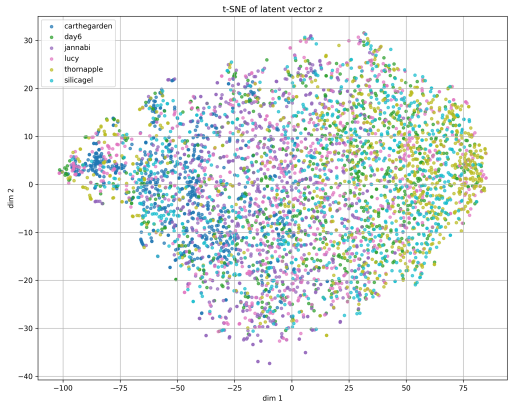
(a) Epoch 0



(b) Epoch 32



(c) Epoch 64



(d) Epoch 80

Figure 2: t-SNE visualization of latent vectors at different epochs 0, 32, 64, 80. Color indicates band label. Clusters become more spread out as training progresses, though band-wise separability remains limited.

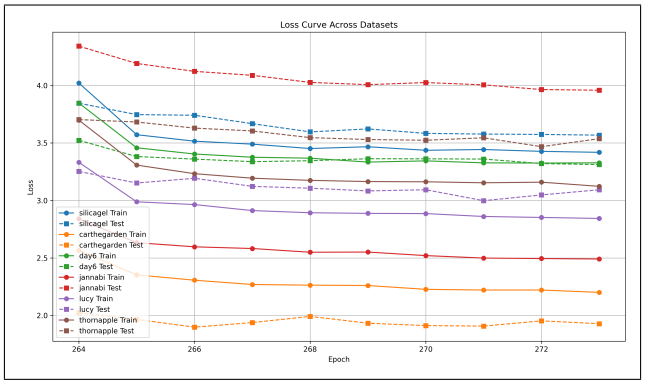


Figure 3: Loss curves for each band's train/test sets during decoder fine-tuning.

diffusion-based decoders, could significantly reduce inference latency without compromising quality.

5. LIMITATIONS

Training was carried out only on instrumental tracks, as the inclusion of voice components introduced instability in convergence. Furthermore, limited computational resources restricted the batch size and total training time. More diverse data and prolonged training are expected to improve model generalization and stylistic fidelity.

6. CONCLUSIONS

We presented a band-to-band style transfer model for long-form music using a universal music translation framework [9]. Our experiments demonstrated successful representation learning in the latent space, as well as genre-specific decoding using band-wise decoders. Although inference speed and stylistic separation remain open challenges, the results indicate a strong potential for further development.

As future work, we aim to incorporate diffusion-based methods such as MusicLDM [3], by overfitting on band-specific prompts and applying AdaIN to diffusion noise for improved quality, generalization and as well as inference speed.

7. AUTHOR CONTRIBUTIONS

"Jooeun Lim" curated the datasets for Lucy, Car, the Garden, Jannabi, and Thornapple. She conducted the full fine-tuning experiments and adapted the training pipeline to support multi-decoder training on a single GPU. She also participated in interpreting the results and analyzing latent space dynamics.

"Taehui Lee" curated the datasets for Day6 and Silicagel. She implemented the decoder-only and post-adaptive fine-tuning strategies. She also developed the visualization tools and generated loss function graphs and latent space plots for each dataset, which were used in the analysis.

8. REFERENCES

[1] Stability AI. Stable audio. `https://stability.ai/research/stable-audio`, 2023.

[2] G. Brunner, Y. Konrad, Y. Wang, and R. Wattenhofer. Symbolic music genre transfer with cyclegan. In *Proc. Int. Society for Music Information Retrieval Conf. (ISMIR)*, pages 747–753, 2018.

[3] K. Chen, H. Liu, J. Xu, Z. Zhang, and M. D. Plumbley. Musicldm: Enhancing novelty in text-to-music generation using beat-synchronous mixup strategies. In *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2024.

[4] C. Donahue, J. McAuley, and M. Puckette. Adversarial audio synthesis. In *Proc. International Conference on Learning Representations (ICLR)*, 2019.

[5] X. Huang and S. Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, pages 1501–1510, 2017.

[6] T. Karras, S. Laine, and T. Aila. A style-based generator architecture for generative adversarial networks. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4401–4410, 2019.

[7] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley. Freesound dataset and baseline for general-purpose audio tagging. In *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pages 626–630, 2020.

[8] H. Liu, Z. Chen, Y. Yang, Q. Kong, Y. Huang, and M. D. Plumbley. Audioldm: Text-to-audio generation with latent diffusion models. *arXiv preprint arXiv:2301.12503*, 2023.

[9] N. Mor, L. Wolf, A. Polyak, and Y. Taigman. A universal music translation network. In *Proc. International Conference on Learning Representations (ICLR)*, 2019.

[10] R. Prenger, R. Valle, and B. Catanzaro. Waveglow: A flow-based generative network for speech synthesis. In *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3617–3621, 2019.

[11] K. Qian, Y. Zhang, S. Chang, X. Yang, and M. Hasegawa-Johnson. Autovc: Zero-shot voice style transfer with only autoencoder loss. In *Proc. International Conference on Machine Learning (ICML)*, pages 5210–5219, 2019.

[12] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022.

[13] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 2016.

[14] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, pages 2223–2232, 2017.