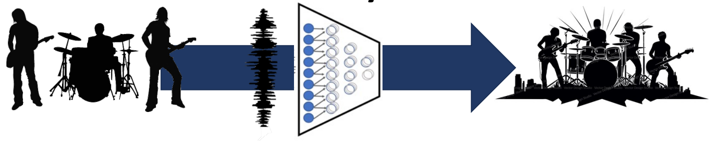
2025 GCT634 Final Project

Band-to-band Style Transfer via Universal Music Translation Network
: Hearing DAY6 as Silicagel

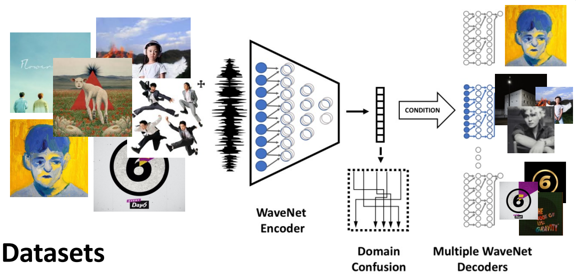
20244418 Jooeun Lim
20253617 Taehui Lee

Introduction

Can we mimic the style of the band?



Method



Datasets

Source: Youtube instrumental playlists

Each dataset was used to train a separate decoder

- Day6: 2h 30m
- Silicagel: 47m
- Carthegarden: 47m
- Jannabi: 1h 4m
- Lucy: 4h 52m
- Thornapple: 3h 16m

https://www.youtube.com/watch?v=HrZwZj99Ja4
https://www.youtube.com/watch?v=QZTvnD6Kznk
https://www.youtube.com/playlist?list=PLydEchhAPYKRdikZwJKtgPnfUl4aUW5sx
https://www.youtube.com/playlist?list=PL0SWBxUZ-emhSrzb-h3QQUYTVcPXJuj6H
https://www.youtube.com/playlist?list=PL3OR2GWIIqPrDwhfI44BgOFoICvBgQCVN
https://www.youtube.com/playlist?list=PLMaFEF_sZ1ZqaBsTZCo1WA0j4-isnWv-N

Experiments

① Pretrained encoder + decoder fine-tuning

- Epoch: 10 / Iteration: 100
- Learning rate: 1e-4
- GPU: RTX 3090 * 1
- Batch size: 16

② Pretrained encoder + full-training

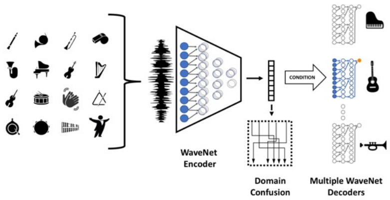
- Epoch: 64 / Iteration: 1000
- Learning rate: encoder 1e-3 / decoder 1e-4
- GPU: RTX 4080 * 2
- Batch size: 8

③ ② + decoder fine-tuning

- Epoch: 10 / Iteration: 100
- Learning rate: 1e-4
- GPU: RTX 3090 * 1
- Batch size: 16

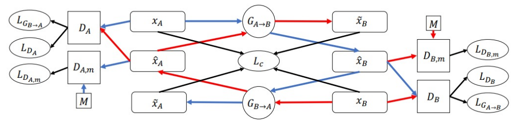
Previous Work

Baseline: A Universal Music Translation Network



Noam Mor, Lior Wolf, Adam Polyak, and Yaniv Taigman, "A Universal Music Translation Network," in *Proc. Int. Conf. on Learning Representations (ICLR)*, 2019.

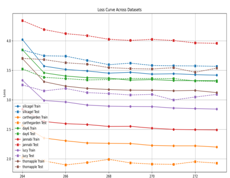
Symbolic Music Genre Transfer with CycleGAN



G. Brunner, Y. Wang, R. Wattenhofer, and S. Zhao, "Symbolic Music Genre Transfer with CycleGAN," in Proc. 30th Int. Conf. on Tools with Artificial Intelligence (ICTAI), Volos, Greece, 2018.

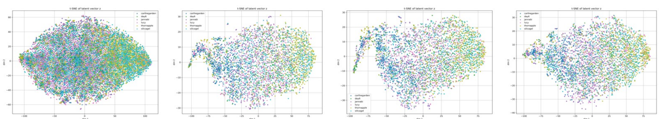
Results

Loss Performance (Exp. ①)



Listen to our Demo!

Latent Vector Distribution (exp ②)

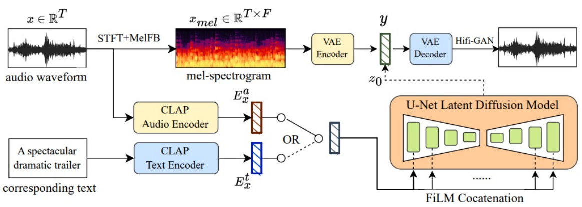


Epoch=0 Epoch=32 Epoch=64 Epoch=80

Further Works

MusicLDM with Noise AdaIN

- Comparison against diffusion-based model
- Finetuning with band labels as text embeddings conditions
- Results will be included in the final report
- Expected advancements
  - Enhanced model capacity
  - Faster inference speed
  - Zero-shot inference capability



Chen, Ke, et al. "Musicldm: Enhancing novelty in text-to-music generation using beat-synchronous mixup strategies." ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2024.

Conclusion

Discussions

1. Latent Vector Distribution
   - The distribution had the tendency to spread as the epoch progressed.
   - However, separation between different bands remains unclear.
   - Regarding that the vector is derived from the clipped data and doesn't have time dependency, the result is rather positive.
2. Loss Performance
   - Loss decreased consistently indicating that training was effective.
   - The difficulty in style discrimination didn't change across time.
   - The model shows potential for further improvement with extended training.
3. Inference time
   - Generating a 5-second sample took approximately 2.9 minutes, which were too long.

Limitations

- Training with voice track was excluded to training difficulty.
- Additional training resources would likely improve model performance and convergence.