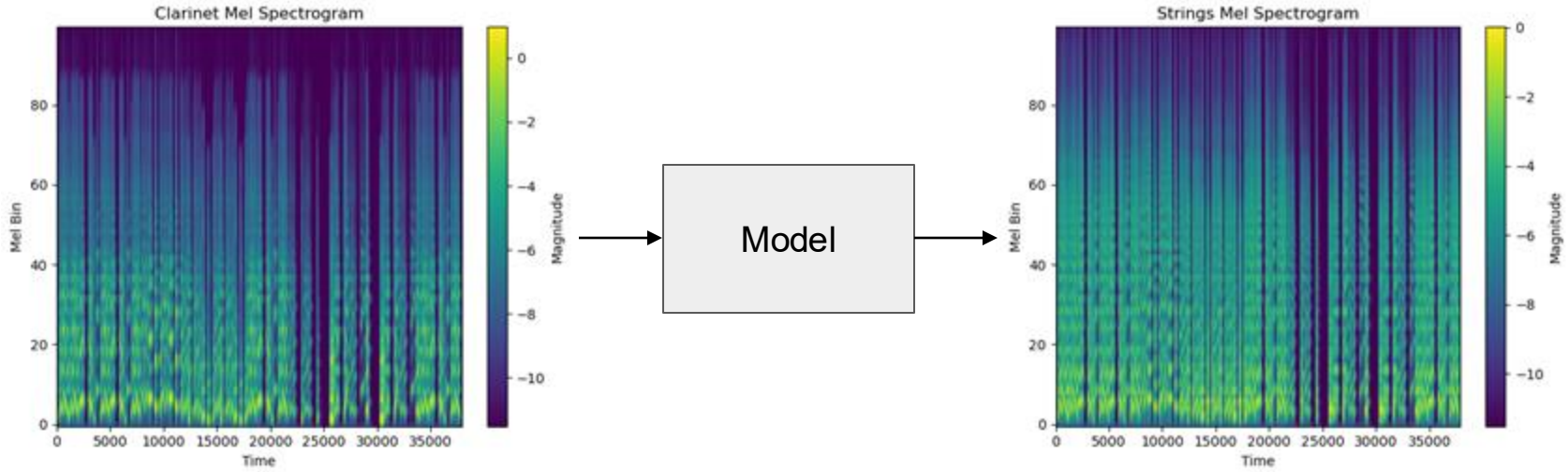
Evaluating Different Input Representations for Timbre Transfer

Ryan Bowering

Goal: Change the timbre of a music recording



The current methods are still immature

Musical timbre style transfer with diffusion model

Hong Huang[1], Junfeng Man[2,3], Luyao Li[1] and Rongke Zeng[1]

[1] School of Computer Science, Hunan University of Technology, Zhuzhou, China
[2] School of Intelligent Manufacturing, Hunan First Normal University, Changsha, China
[3] Key Laboratory of Industrial Equipment Intelligent Perception and Maintenance in College of Hunan Province, Hunan First Normal University, Changsha, China

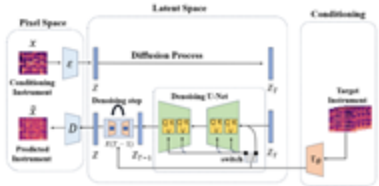
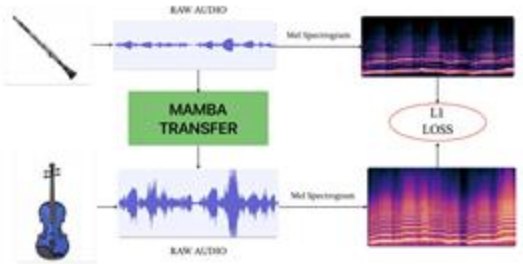


Figure 2 **Models of timbre transfer.** Full-size DOI: 10.7717/peerj-cs.2194/fig-2

MambaTransfer: raw audio musical timbre tranfer using selective state-space models

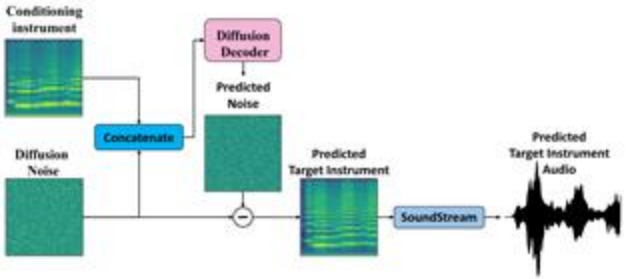
Tesi di Laurea Magistrale in
Music and Acoustic Engineering - Ingegneria Musicale ed Acustica

Guglielmo Fratticioli, 10821800



TIMBRE TRANSFER USING IMAGE-TO-IMAGE DENOISING DIFFUSION IMPLICIT MODELS

Luca Comanducci Fabio Antonacci Augusto Sarti
Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, Italy
luca.comanducci@polimi.it, fabio.antonacci@polimi.it, augusto.sarti@polimi.it



This project attempts to organize the current approach

- Flow matching with DiT backbone
- Compare performance of mel-spec, CQT, and raw audio

Change of Scope

- There aren’t really any CQT to waveform vocoders
- The github for the CQT model is not available
- Training one seemed out of scope

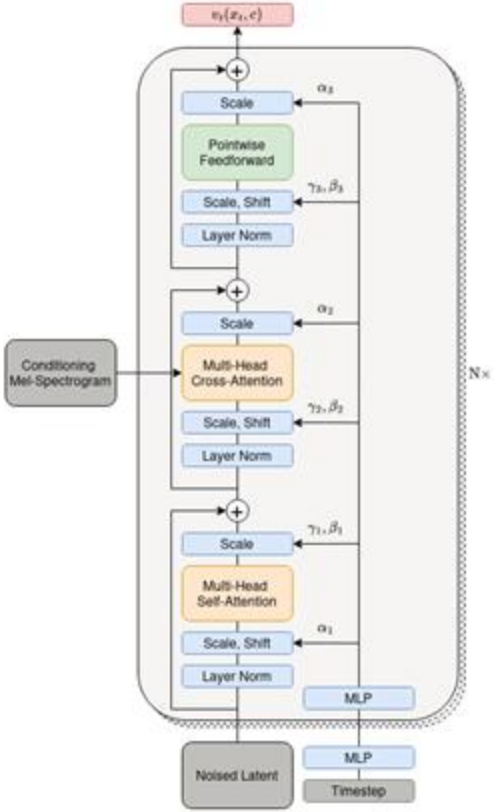
=> I only investigated the mel spectrogram and raw waveform

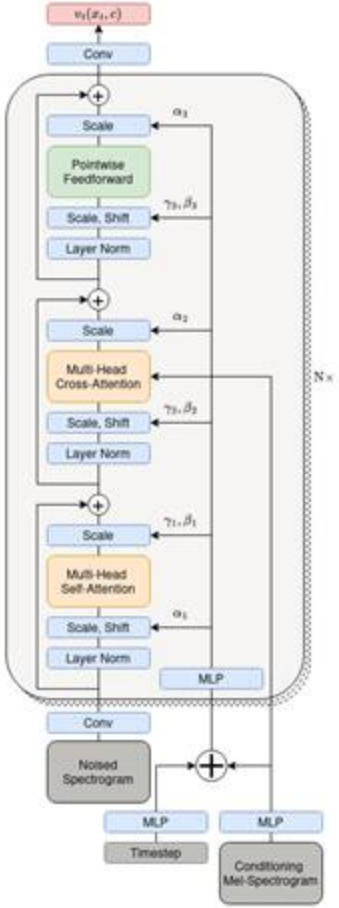
The StarNet dataset is ideal for this project

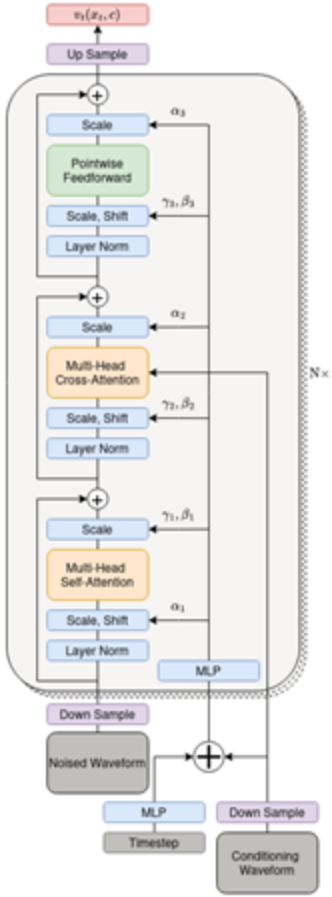
Contains recordings of different instrument pairs playing the same piece

1. ~~**001.0.wav:** the clarinet vibraphone mixture~~
2. **001.1.wav:** the clarinet track
3. **001.2.wav:** the vibraphone track
4. ~~**001.3.wav:** the strings piano mixture~~
5. **001.4.wav:** the strings track
6. **001.5.wav:** the piano track

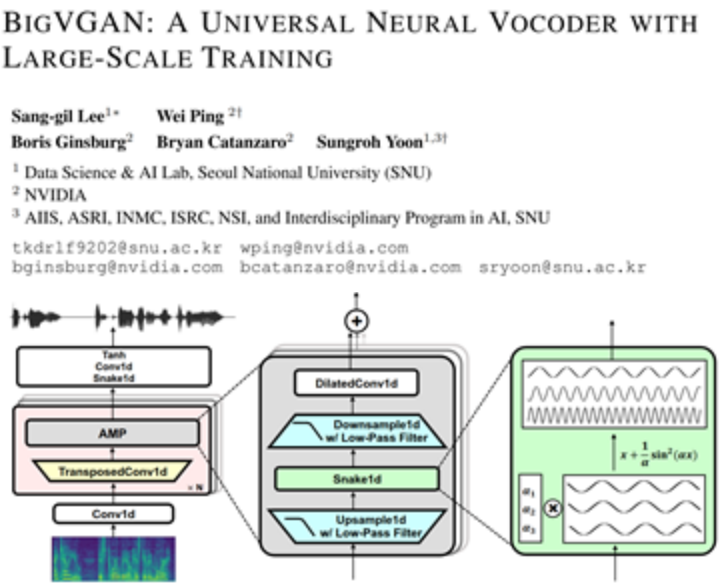
Model Architecture







Mel Spectrogram Vocoder



BIGVGAN: A UNIVERSAL NEURAL VOCODER WITH LARGE-SCALE TRAINING

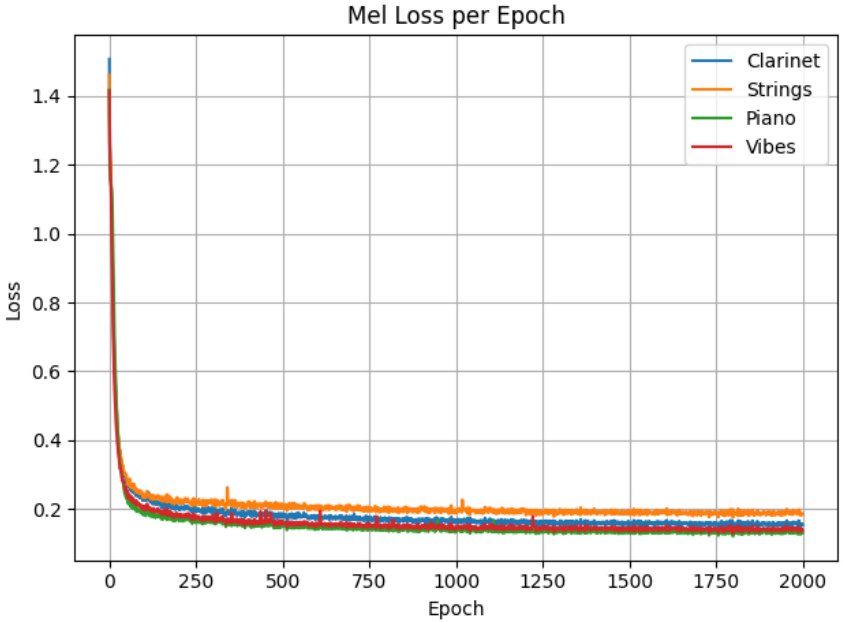
Sang-gil Lee[1*] Wei Ping[2†]
Boris Ginsburg[2] Bryan Catanzaro[2] Sungroh Yoon[1,3†]

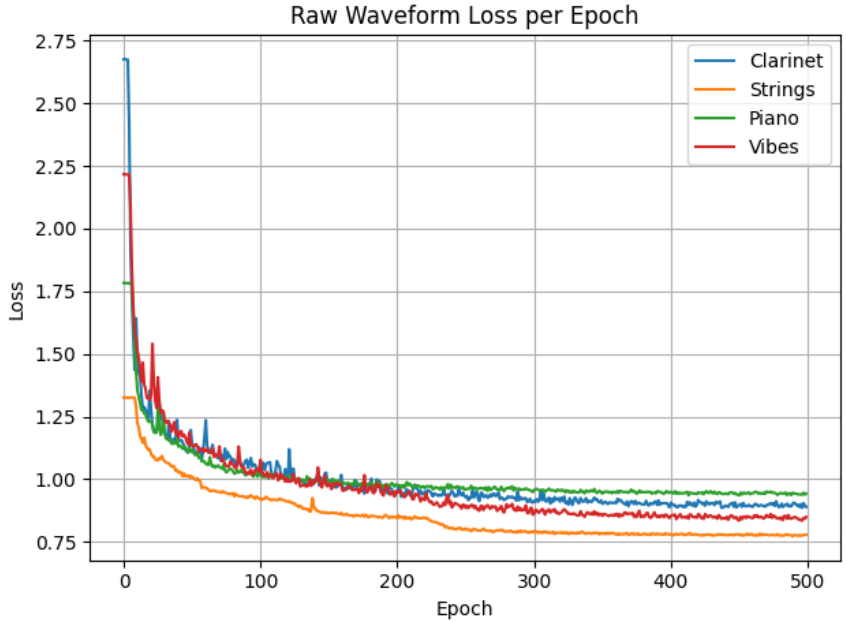
[1] Data Science & AI Lab, Seoul National University (SNU)
[2] NVIDIA
[3] AIIS, ASRI, INMC, ISRC, NSI, and Interdisciplinary Program in AI, SNU

tkdrlf9202@snu.ac.kr wping@nvidia.com
bginsburg@nvidia.com bcatanzaro@nvidia.com sryoon@snu.ac.kr

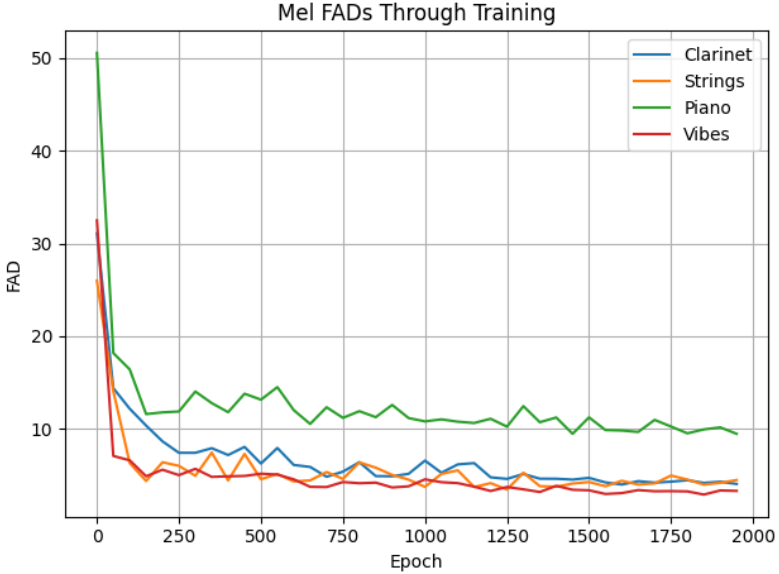
- The mel spectrograms were calculated at 24 kHz sampling rate
- Reconstructed audio was resampled to 16 kHz for FAD
- The raw waveform model worked with 16 kHz data to minimize computational costs

Results


Mel Loss per Epoch
Clarinet
Strings
Piano
Vibes
Loss
Epoch
0
250
500
750
1000
1250
1500
1750
2000
0.2
0.4
0.6
0.8
1.0
1.2
1.4


Raw Waveform Loss per Epoch
Clarinet
Strings
Piano
Vibes
Loss
Epoch
0
100
200
300
400
500
0.75
1.00
1.25
1.50
1.75
2.00
2.25
2.50
2.75

Results


Mel FADs Through Training
Clarinet
Strings
Piano
Vibes
FAD
Epoch
50
40
30
20
10
0
250
500
750
1000
1250
1500
1750
2000


Raw Waveform FADs Through Training
Clarinet
Strings
Piano
Vibes
FAD
Epoch
80
60
40
20
0
100
200
300
400
500

Results

FAD Comparison

	Strings to Clarinet	Clarinet to Strings	Vibes to Piano	Piano to Vibes
Mel	4.06	4.49	9.47	3.31
TS	31.82	15.07	25.58	14.52

Table 5 Results of the objective evaluation contrasted with baseline models.

Model	Task					
	Piano to Guitar		Piano to Vibraphone		Vibraphone/Clarinet to Piano/Strings	
	FAD	JD	FAD	JD	FAD	JD
VAE-GAN	8.41	0.54	9.16	0.56	12.52	0.67
Music-Star	6.47	0.39	7.43	0.41	10.93	0.57
DiffTransfer	3.34	0.31	4.56	0.28	6.73	0.46
DiffTransfer (DiffWave)	3.20	0.30	4.31	0.28	6.43	0.47
ours	3.16	0.32	4.22	0.29	6.37	0.48

Examples

Input Type	Conditioning	Target	Model Output
Mel Spectrogram			
Mel Spectrogram			
Mel Spectrogram			
Waveform			
Waveform			
Waveform			

Future Work

- Latent diffusion for raw audio, and more computational resources
- Mix and match input/conditioning signals (i.e. audio input with mel conditioning)
- Train a CQT vocoder to try a CQT-based model
- Unpaired timbre transfer