TimbreTron: A WaveNet(CycleGAN(CQT(Audio))) Pipeline for Musical Timbre Transfer

Sicong Huang[1,2] Qiyang Li[1,2] Cem Anil[1,2] Xuchan Bao[1,2] Sageev Oore[2,3] Roger B. Grosse[1,2]

[1] University of Toronto [2] Vector Institute [3] Dalhousie University

Introduction

- **Musical Timbre Transfer** refers to manipulating the timbre of a sound sample from one instrument to match another instrument while preserving other musical content such as pitch and volume.
- **Modeling timbre** is very hard and is important for musicians. Timbre is often referred to as the "psycho-acoustician waste basket".
- **TimbreTron** is a pipeline that uses CQT, CycleGAN, and WaveNet to perform timbre transfer with high-quality output
- **Check out our Project Page and Video!** *https://www.cs.toronto.edu/~huang/TimbreTron/index.html*

Background

- **Time Frequency Analysis** refers to techniques that aim to measure how the signal's frequency domain representation changes over time.
- **Short Time Fourier Transform (STFT)** and **Constant-Q-Transform (CQT)** are examples of the time frequency analysis techniques.

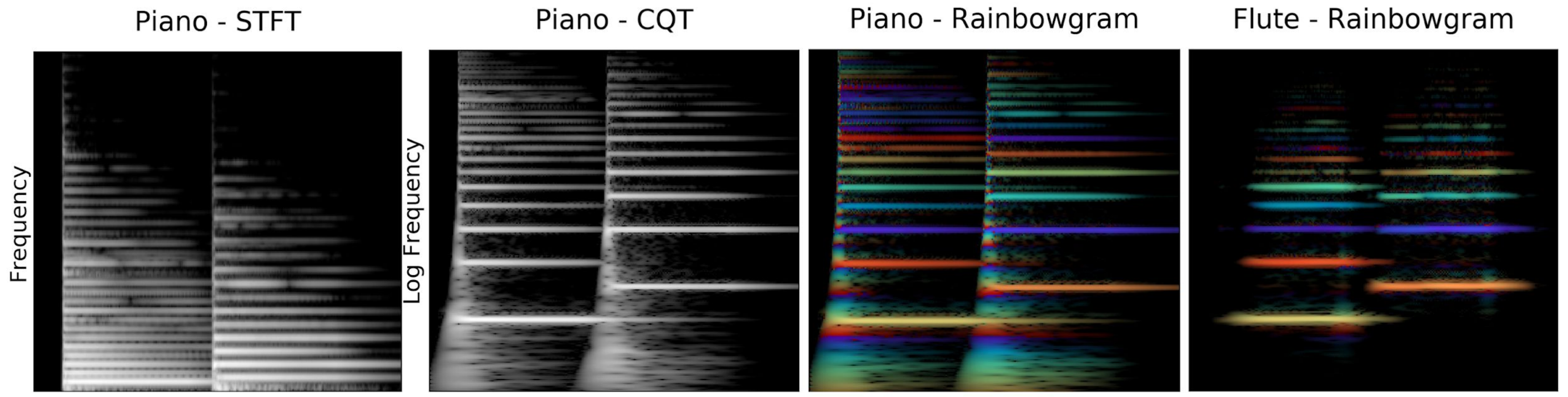

Figure 1. CQT, STFT, and their rainbowgrams. Rainbowgrams are CQT spectrograms with magnitude represented by intensity and instantaneous frequency by color

- **WaveNet [1]** is an auto-regressive generative model for generating raw audio waveform with high quality.
- **CycleGAN [2]** is a method for unsupervised domain transfer: learning a mapping between two domains without any paired data.

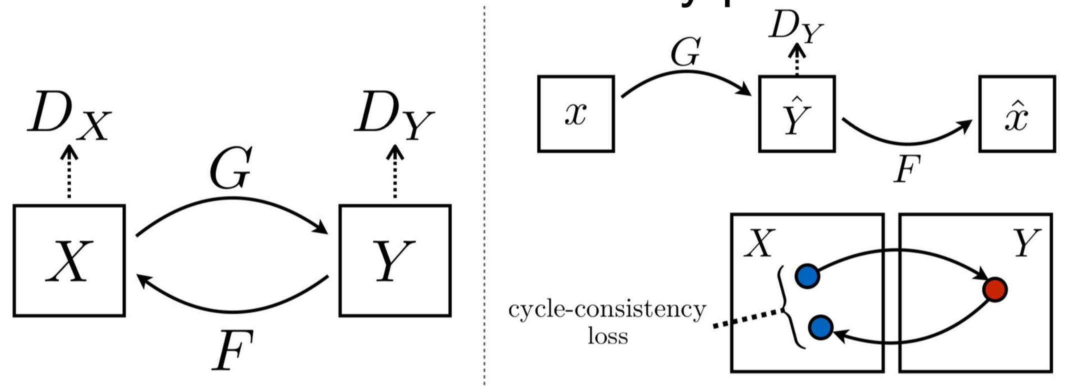

Figure 2. CycleGAN model for unsupervised domain transfer.

Method

TimbreTron Pipeline

Step 1: Convert source audio to its CQT representation in the source domain.
Step 2: Translate the CQT spectrogram to the target domain.
Step 3: Reconstruct the target domain audio from the target spectrogram using a conditional waveNet synthesizer that is trained on the target domain

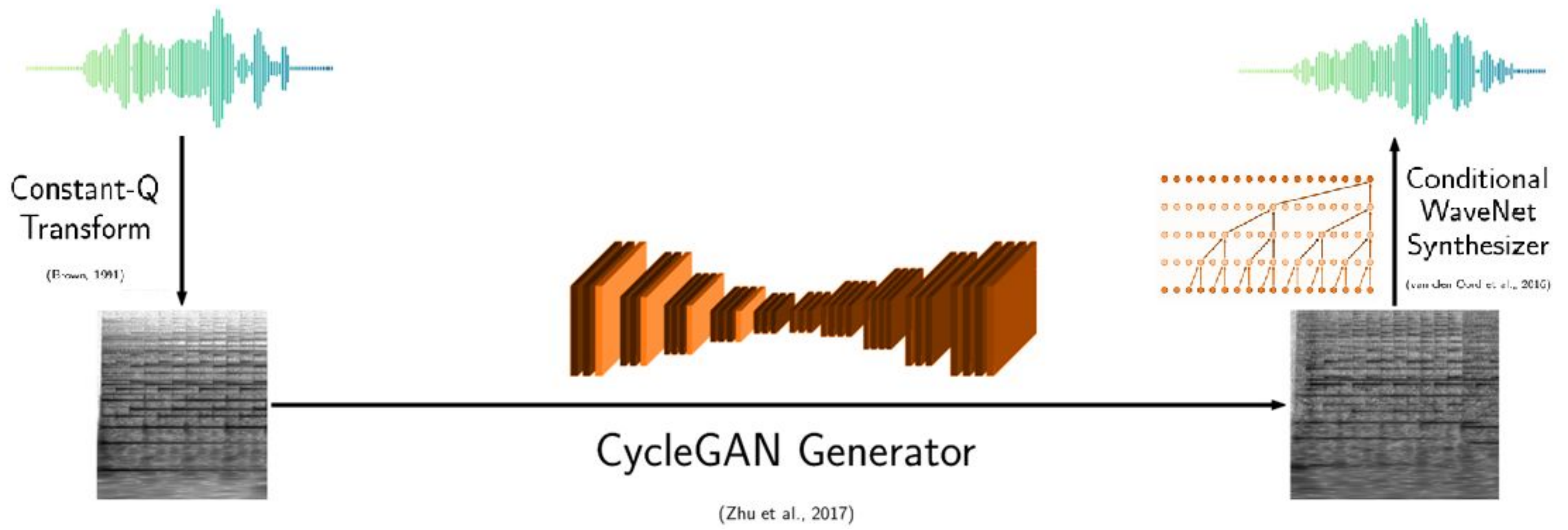

Figure 3. TimbreTron pipeline overview

- **Why CQT?**
  - Unlike STFT, CQT has higher frequency resolution towards lower frequencies (for lower register instruments) and higher time resolution towards higher frequencies (for fine timing of rhythms).
  - Convolution on CQT spectrograms is equivariant under pitch shift. (Figure 1)
  - Pitch can be manipulated independent of timbre and rhythm by shifting the CQT spectrogram vertically (see examples in the video).
- **Why CycleGAN?**
  - CycleGAN performs high-quality unpaired image-to-image translation.
  - Spectrogram translation problem between two instrument domains can be treated as a instance of unsupervised image translation.
- **Why WaveNet?**
  - CQT cannot be easily inverted to recover a waveform. WaveNet can be easily adapted to generate high quality audio from low-level acoustic features (e.g., spectrograms) [3].
  - Tempo can be manipulated by sub/over-sampling spectrogram windows per time step into the WaveNet (see examples in the video).

Results

- Ablation study demonstrates that every architectural change was necessary
- Evaluation with Amazon Mechanical Turk (AMT) confirmed that:
  - TimbreTron can transfer timbre while preserving musical content.
  - CQT works better than STFT.

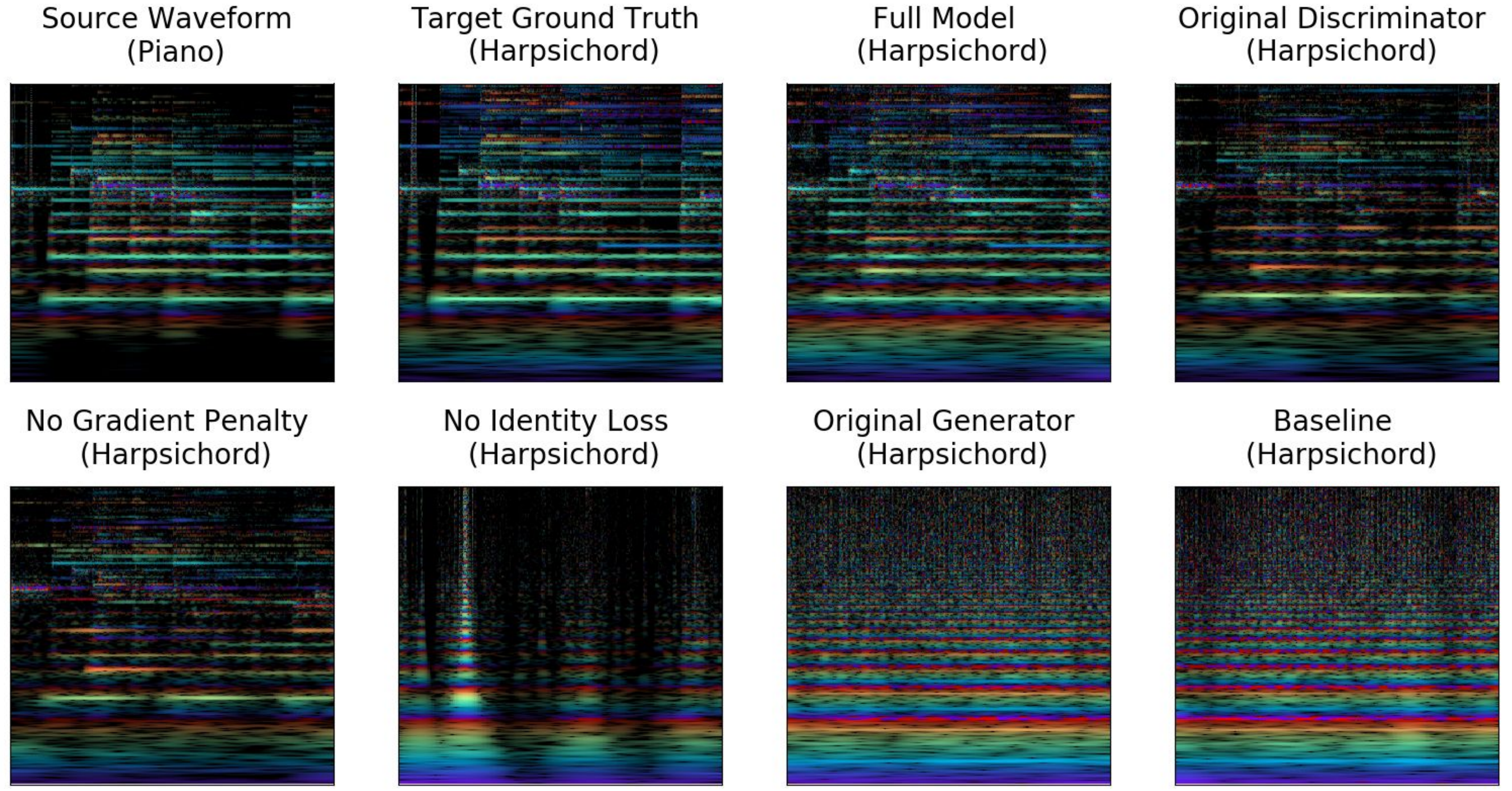

Figure 4. Ablation Study

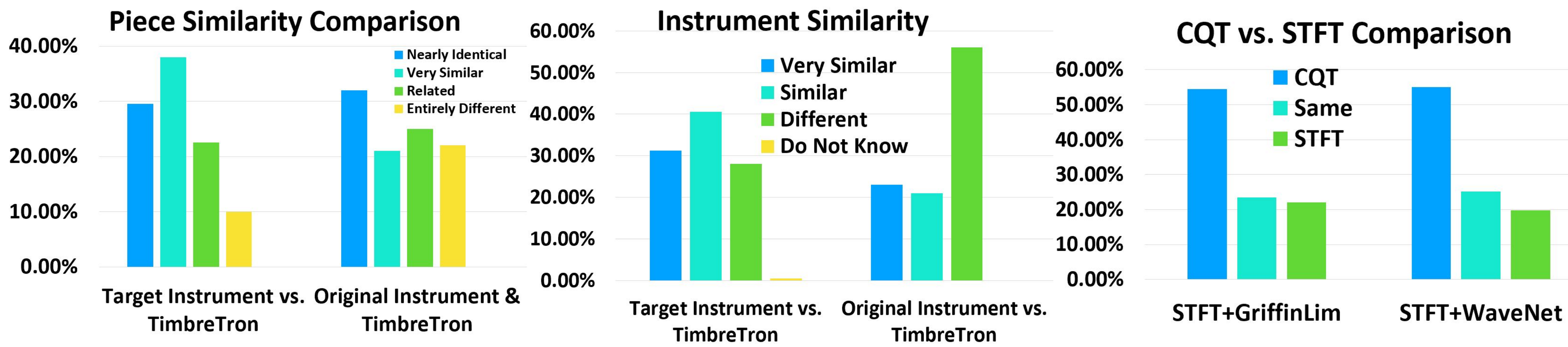

Figure 5. AMT Results

[1] Aäron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew W. Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. CoRR, abs/1609.03499, 2016.
[2] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. CoRR, abs/1703.10593, 2017.
[3] Jonathan Shen, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, RJ Skerry-Ryan, et al. Natural TTS synthesis by conditioning wavenet on mel spectrogram predictions. arXiv preprint arXiv:1712.05884, 2017.