DrumGAN: Synthesis of Drum Sounds With Timbral Feature Conditioning Using GANs

Javier Nistal, Stefan Lattner, and Gaël Richard

Sony CSL · TELECOM Paris · IP PARIS · MIP Frontiers

Introduction

Audio synthesizers have **complicated parameters** with little **perceptual correspondence nor musical meaning**. Also, the type of **sounds** they can produce **are limited by the synthesis method** (e.g. additive, subtractive).

DrumGAN is a Progressive Growing GAN (PGAN) that can synthesize a wide variety of drum sounds and that enables steering the synthesis according to parameters that respond to human perception.

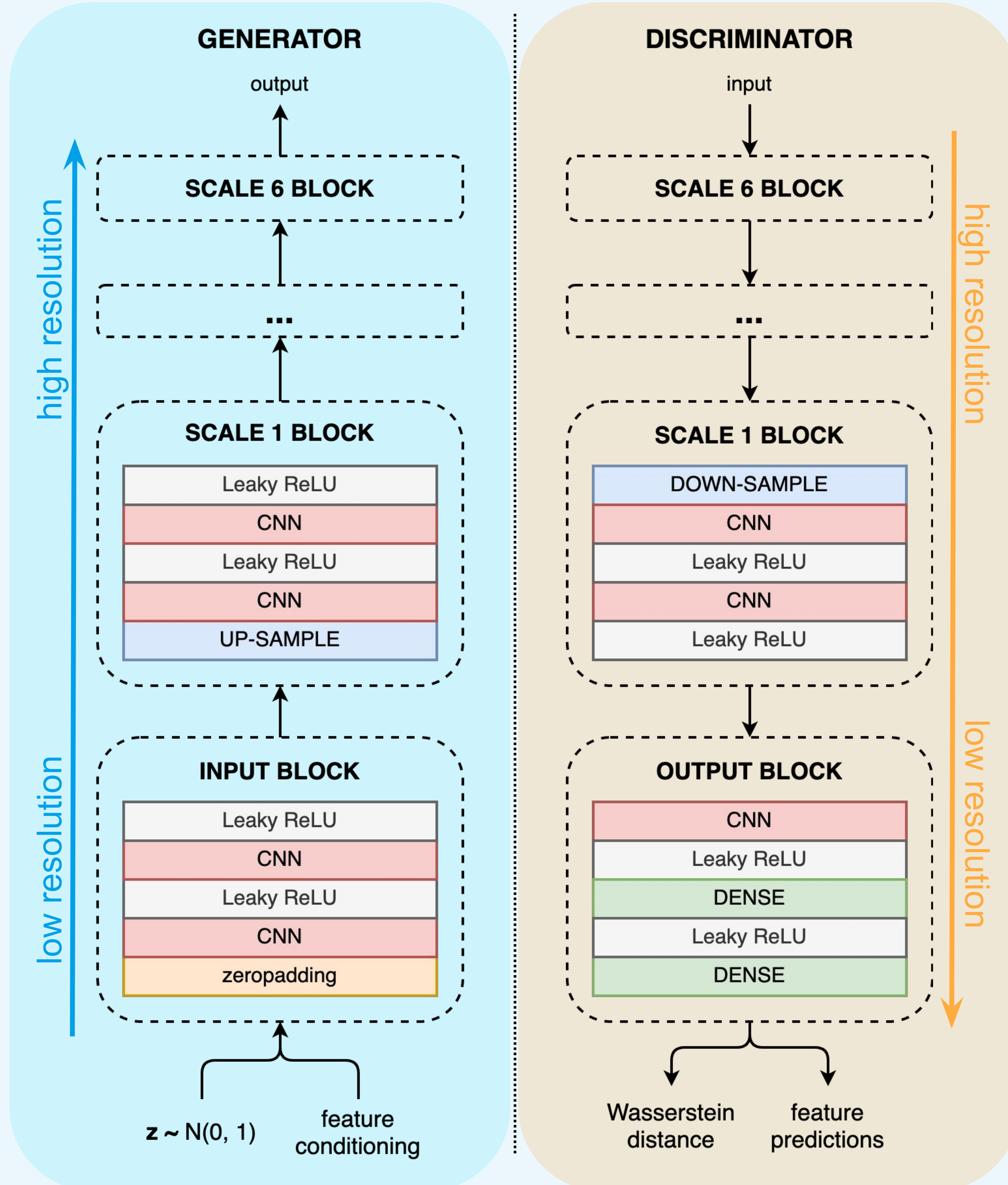
Dataset

- ~**300k** one-shot, 1s-long and aligned audio samples
- **Kicks (K), Snares (S) and Cymbals (C)** classes
- **16kHz sampling-rate**
- **90/10%** train-validation split
- **Complex STFT** representation
  - window size: 2048
  - hop size: 512

Audio-Commons Features

- Audio Commons perceptual models → high-level timbral features of the sound
- Human ratings given to sounds from Freesound
- Linear regression models of spectral and temporal low-level features (e.g., spectral centroid, dynamic-range)
- All features are normalised to the range [0-1]

Architecture & Training Procedure



Progressive Growing GAN

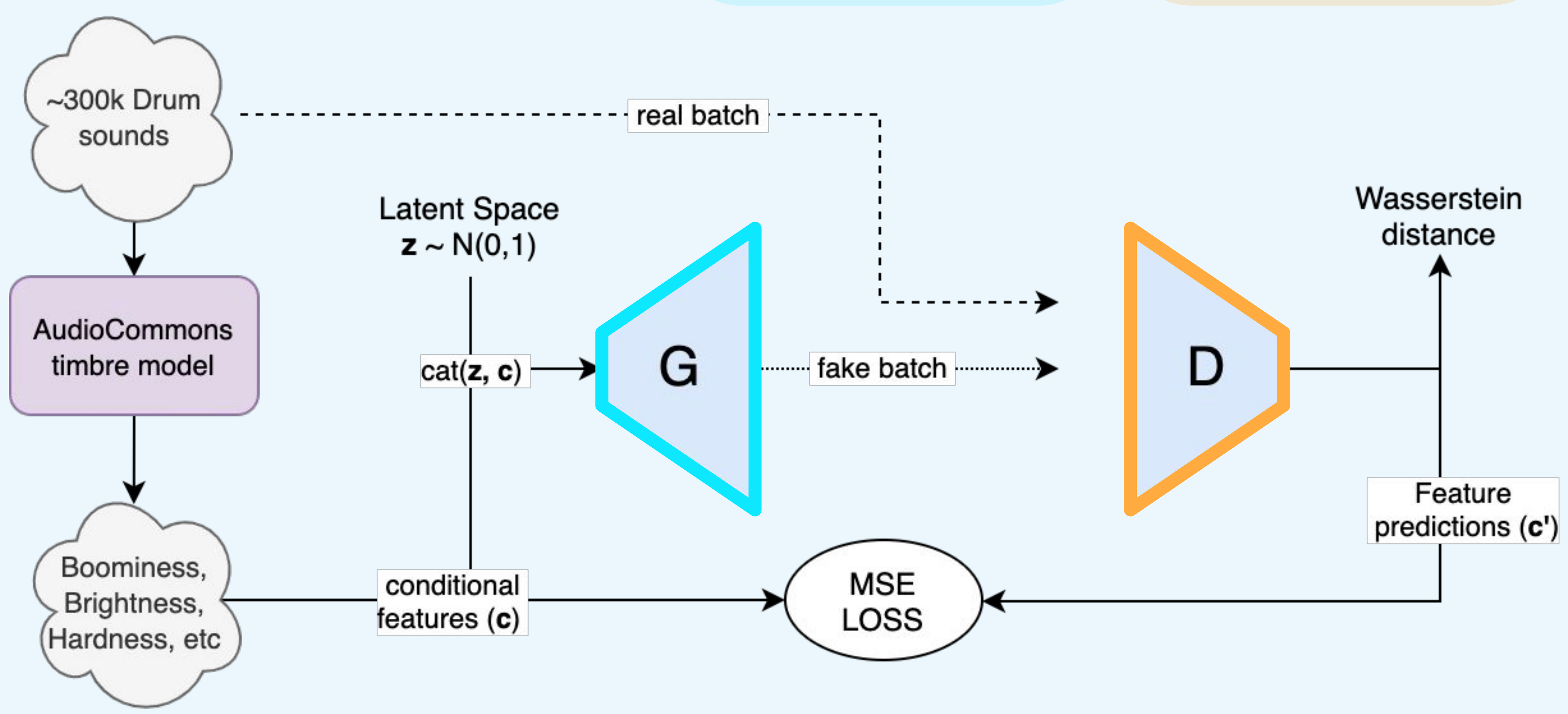
Input Block:
[**z**, **c**] → [batch, ch, f_{s0}, t_{s0}]

Scale blocks: progressively added to the network while training. CNN out channels for each scale:

[256, 128, 128, 128, 128, 64]

Training:

- 1.1M iterations (~200k i/scale)
- batch-size: [30, 30, 20, 20, 12, 12]
- Adam optimizer
- learning rate: 1e-3.

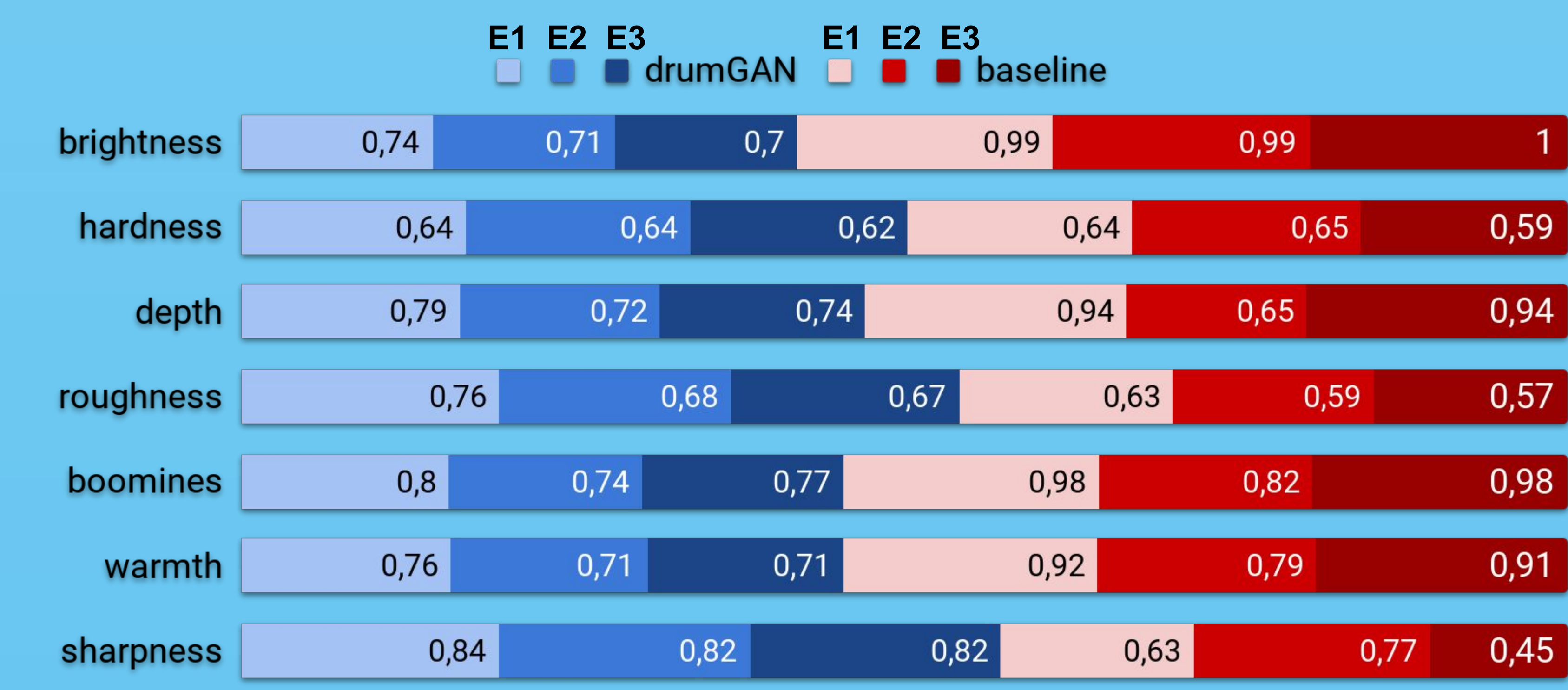


Results

Attribute coherence

A specific feature f_i is set to 0.2, 0.5, and 0.8, keeping the other features and z fixed. The outputs of **G** ($fx_{i_{0.2/0.5/0.8}}$) are evaluated with the Audio Commons Models. Three conditions are examined:

E1: $fx_{i_{0.2}} < fx_{i_{0.5}}$ E2: $fx_{i_{0.5}} < fx_{i_{0.8}}$ E3: $fx_{i_{0.2}} < fx_{i_{0.5}}$



DISCUSSION

baseline yields high accuracies for features describing the global frequency distribution (e.g., brightness, depth). *drumGAN* performs better on features describing complex frequential relationships (e.g., roughness, sharpness).

Scores & Distances

Inception Score (IS), **Kernel Inception Distance** (KID), **Fréchet Audio Distance** (FAD)

→ *real data*: scores and distances on real data
→ *uncond:* unconditional drumGAN
→ *drumGAN_train*: conditioned on training labels
→ *drumGAN_val*: conditioned on validation labels
→ *drumGAN_rand*: conditioned on random labels
→ *baseline*: UNet conditioned on real labels

DISCUSSION

The **IS** of *drumGAN* is **close to that of real data** → outputs are assignable into {**K, S, C**}
uncond yields worse KID and FAD → the **features help generating more realistic samples**.
drumGAN* outperforms the *baseline *metrics for all conditional settings*.

SOUND EXAMPLES ➡ **PAPER** ➡ ➡