GANSYNTH: ADVERSARIAL NEURAL AUDIO SYNTHESIS

ABSTRACT
摘要

Efficient audio synthesis is an inherently difficult machine learning task, as human perception is sensitive to both global structure and fine-scale waveform coherence. Autoregressive models, such as WaveNet, model local structure but have slow iterative sampling and lack global latent structure. In contrast, Generative Adversarial Networks (GANs) have global latent conditioning and efficient parallel sampling, but struggle to generate locally-coherent audio waveforms. Herein, we demonstrate that GANs can in fact generate high-fidelity and locally-coherent audio by modeling log magnitudes and instantaneous frequencies with sufficient frequency resolution in the spectral domain. Through extensive empirical investigations on the NSynth dataset, we demonstrate that GANs are able to outperform strong WaveNet baselines on automated and human evaluation metrics, and efficiently generate audio several orders of magnitude faster than their autoregressive counterparts.

高效的音频合成是天生就是一个困难的机器学习任务，因为人类的感知系统对全局结构和细尺度波形一致性都十分敏感。自回归模型（如 WaveNet）对局部结构建模但迭代采样速度慢且缺少全局的潜在结构。相比之下，对抗生成网络（GAN）具有全局的潜在制约和高效的并行采样，但难以生成局部一致的音频波形。在这里，我们论证了 GAN 可以通过在频谱域中对具有足够频率分辨率的 log magnitudes 和 instantaneous frequencies 进行建模来生成高保真和局部一致的音频。通过对 NSynth 数据集的测试，我们证明了 GAN 能够在自动和人工评估方面优于 WaveNet，并且比这些自回归的模型的生成音频速度快多个数量级。

1 INTRODUCTION
1 简介

Neural audio synthesis, training generative models to efficiently produce audio with both high-fidelity and global structure, is a challenging open problem as it requires modeling temporal scales over at least five orders of magnitude (0.1ms to 100s). Large advances in the state-of-the art have been pioneered almost exclusively by autoregressive models, such as WaveNet, which solve the scale problem by focusing on the finest scale possible (a single audio sample) and rely upon external conditioning signals for global structure (van den Oord et al., 2016). This comes at the cost of slow sampling speed, since they rely on inefficient ancestral sampling to generate waveforms one audio sample at a time. Due to their high quality, a lot of research has gone into speeding up generation, but the methods introduce significant overhead such as training a secondary student network or writing highly customized low-level kernels (van den Oord et al., 2018; Paine et al., 2016). Furthermore, since these large models operate at a fine timescale, their autoencoder variants are restricted to only modeling local latent structure due to memory constraints (Engel et al., 2017).

神经网络音频合成，训练生成模型以有效地生成拥有高保真度和全局结构的音频是一个具有挑战性的开放问题，因为它需要对至少五个数量级（0.1ms 到 100s）的时间尺度进行建模。在最先进的模型中，进展几乎来自于自回归模型，如

WaveNet，它通过关注尽可能最小的尺度（单个音频样本）和依靠外部条件信号实现全局结构来解决尺度问题。这样做的代价是采样速度缓慢，因为他们需要一步一步的通过低效的之前采样来生成后面的音频样本。由于他们生成音频的质量高，有很多学者研究这个模型如何加速生成，但这些方法引入了大量的开销，如训练一个子网络或编写高度定制的低级内核。此外，由于这些大型模型在精细的时间尺度下运作，因此其自动编码器变异是受限制的，因为内存的原因，只能建模局部潜在结构。

On the other end of the spectrum, Generative Adversarial Networks (GANs) (Goodfellow et al., 2014) have seen great recent success at generating high resolution images (Radford et al., 2016; Arjovsky et al., 2017; Gulrajani et al., 2017; Berthelot et al., 2017; Kodali et al., 2017; Karras et al., 2018a; Miyato et al., 2018). Typical GANs achieve both efficient parallel sampling and global latent control by conditioning a stack of transposed convolutions on a latent vector, The potential for audio GANs extends further, as adversarial costs have unlocked intriguing domain transformations for images that could possibly have analogues in audio (Isola et al., 2017; Zhu et al., 2017; Wolf et al., 2017; Jin et al., 2017). However, attempts to adapt image GAN architectures to generate waveforms in a straightforward manner (Donahue et al., 2019) fail to reach the same level of perceptual fidelity as their image counterparts.
另一方面，对抗生成网络（GANs）最近在生成高分辨率图像方面取得了巨大成功。典型的 GAN 把潜在向量送入多个转置卷积来实现高效的并行采样和全局潜在控制，音频 GAN 的潜力有待进一步扩展，因为对抗性成本为可能和图像具有相同模式的音频开启了有趣的知识迁移。但是，以简单的方式调整图像 GAN 体系结构以生成波形并未能达到与其图像模型相同的水平。

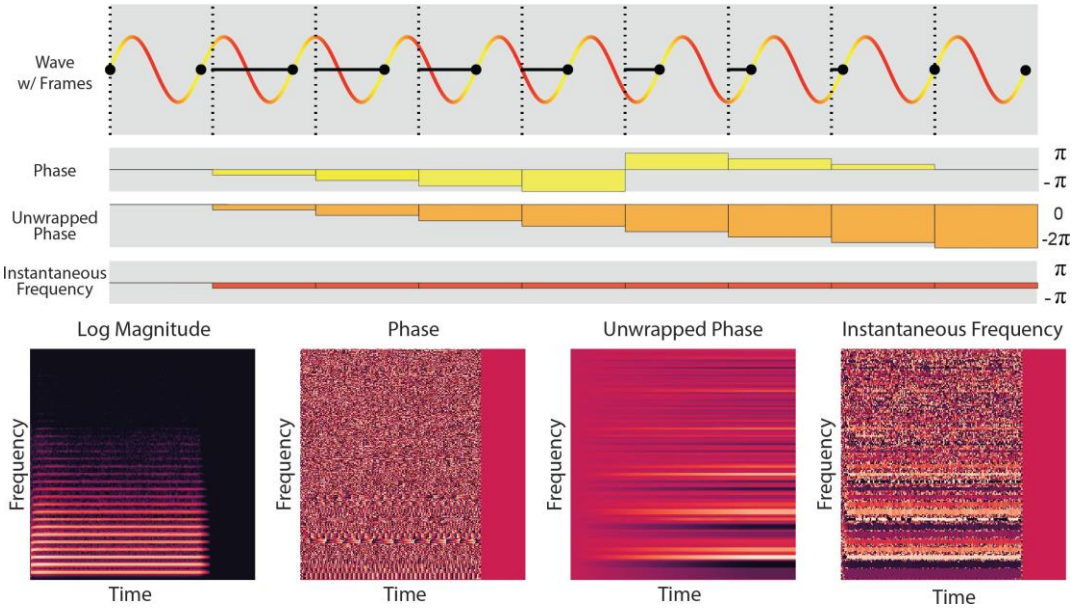

Figure 1: Frame-based estimation of audio waveforms. Much of sound is made up of locally- coherent waves with a local periodicity, pictured as the red-yellow sinusoid with black dots at the start of each cycle. Frame-based techniques, whether they be transposed convolutions or STFTs, have a given frame size and stride, here depicted as equal with boundaries at the dotted lines. The alignment between the two (phase,

indicated by the solid black line and yellow boxes), precesses in time since the periodicity of the audio and the output stride are not exactly the same. Transposed convolutional filters thus have the difficult task of covering all the necessary frequencies and all possible phase alignments to preserve phase coherence. For an STFT, we can unwrap the phase over the 2π boundary (orange boxes) and take its derivative to get the instantaneous radial frequency (red boxes), which expresses the constant relationship between audio frequency and frame frequency. The spectra are shown for an example trumpet note from the NSynth dataset.

图 1：基于帧的音频波形估计。大部分声音是由局部相关的波组成的，具有局部周期性，绘制成红黄色正弦图形，黑点代表每个周期的开始。基于帧的技术，无论是转置卷积还是 STFT，都有给定的帧大小和步长，此处被表示为与虚线的边界距离。由实黑线和黄色框显示的相位之间的对齐，由于音频的周期性和输出步幅不完全相同，需要及时处理。因此，转置卷积滤波器的任务十分艰巨，即覆盖所有必要的频率和所有可能的相位对齐，以保持相位的一致性。对于 STFT，我们可以在 2π 边界（橙色框）的相位上解包，并取其导数来获得瞬时径向频率（红色框），从而表示音频频率和帧频率之间的恒定关系。作为一个示例，对于 NSynth 数据集中的小号音符，显示其频谱。

1.1 GENERATING INSTRUMENT TIMBRES

1.1 生成乐器音色

GAN researchers have made rapid progress in image modeling by evaluating models on focused datasets with limited degrees of freedom, and gradually stepping up to less constrained domains. For example, the popular CelebA dataset (Liu et al., 2015) is restricted to faces that have been centered and cropped, removing variance in posture and pose, and providing a common reference for qualitative improvements (Radford et al., 2016; Karras et al., 2018a) in generating realistic texture and fine-scale features. Later models then built on that foundation to generalize to broader domains (Karras et al., 2018b; Brock et al., 2019).

GAN 的研究人员通过在特定的数据集上评估模型，并逐步进入受限较少的数据集，在图像建模方面已经取得了快速进展。例如，流行的 CelebA 数据集中仅包含已居中和裁剪的脸部，消除了姿势的差异，并为质量改进（生成逼真的纹理和精细尺度特征）提供了参考。后来的模型在此基础上建立起来，涵盖了更广泛的数据。

The NSynth dataset (Engel et al., 2017)2 was introduced with similar motivation for audio. Rather than containing all types of audio, NSynth consists solely of individual notes from musical instruments across a range of pitches, timbres, and volumes. Similar to CelebA, all the data is aligned and cropped to reduce variance and focus on fine-scale details, which in audio corresponds to timbre and fidelity. Further, each note is also accompanied by an array of attribute labels to enable exploring conditional generation.

音频领域，NSynth 数据集的引入具有类似的动机。NSynth 没有包含所有类型的音频，而是仅包含各种音高、音色和音量的乐器的单个音符。与 CelebA 类似，所有数据都对齐和裁剪过，以减少方差并专注于精细尺度的细节，在音频中对应于音色和保真度。此外，每个音符还附带一系列属性标签，以便探索条件生成。

The original NSynth paper introduced both autoregressive WaveNet autoencoders and bottleneck spectrogram autoencoders, but without the ability to unconditionally sample from a prior. Follow up work has explored diverse approaches including frame-based regression models (Defossez et al., 2018), inverse scattering networks (Andreux & Mallat, 2018), VAEs with perceptual priors (Esling et al., 2018), and adversarial regularization for domain transfer (Mor et al., 2019). This work builds on these efforts by introducing adversarial training and exploring effective representations for non-causal convolutional generation as typical found in GANs.
原始的 NSynth 论文介绍了自回归 WaveNet 自动编码器和瓶颈频谱（bottleneck spectrogram）自动编码器，但无法无条件地从之前的样本来采样。后续工作探索了多种方法，包括基于帧的回归模型、反向散射网络、具有感知先验的 VAEs 和领域迁移的对抗性正则化。本文以这些工作为基础，引入对抗性训练，探索非因果卷积生成的有效表示形式，这在 GAN 中是常见的。

1.2 EFFECTIVE AUDIO REPRESENTATIONS FOR GANS

1.2 GAN 的有效音频表示

Unlike images, most audio waveforms—such as speech and music—are highly periodic. Convolutional filters trained for different tasks on this data commonly learn to form logarithmically-scaled frequency selective filter banks spanning the range of human hearing (Dieleman & Schrauwen, 2014; Zhu et al., 2016). Human perception is also highly sensitive to discontinuities and irregularities in periodic waveforms, so maintaining the regularity of periodic signals over short to intermediate timescales (1ms - 100ms) is crucial. Figure 1 shows that when the stride of the frames does not exactly equal a waveform's periodicity, the alignment (phase) of the two precesses over time.
与图像不同，大多数音频波形（如语音和音乐）都是高度周期性的。针对这些数据的不同任务训练的卷积滤波器通常学习形成跨越人类听觉范围的对数频率的选择性滤波器组。人类的感知对周期性波形中的不连续性和不规则性也高度敏感，因此在短到中时间尺度（1ms - 100ms）保持周期性信号的规律性至关重要。图 1 显示，当帧的步长不完全等于波形的周期时，两个对齐（相位）会随时间按岁差进动。
This condition is assured as at any time there are typically many different frequencies in a given signal. This is a challenge for a synthesis network, as it must learn all the appropriate frequency and phase combinations and activate them in just the right combination to produce a coherent waveform. This phase precession is exactly the same phenomena observed with a short-time Fourier transform (STFT), which is composed of strided filterbanks just like convolutional networks. Phase precession also occurs in situations where filterbanks overlap (window or kernel size < stride).
这种情况是确定的，因为在任何时候，在给定的信号中通常有很多不同的频率。这对生成网络来说是一个挑战，因为它必须学习所有适当的频率和相位组合，并以正确的组合激活它们，产生一致性的波形。此相位衰退与短时间的 Fourier 变换（short-time Fourier transform , STFT）观察到的现象完全相同，该变换由 strided 滤波器组组成，就像卷积网络一样。在滤波器组重叠（窗口大小<步长）的情况下，也会发生相位衰退。
In the middle of Figure 1, we diagram another approach to generating coherent

waveforms loosely inspired by the phase vocoder (Dolson, 1986). A pure tone produces a phase that precesses. Un- wrapping the phase, by adding 2π whenever it crosses a phase discontinuity, causes the precessing phase to grow linearly. We then observe that the derivative of the unwrapped phase with respect to time remains constant and is equal to the angular difference between the frame stride and signal periodicity. This is commonly referred to as the instantaneous angular frequency, and is a time varying measure of the true signal oscillation. With a slight abuse of terminology, we will simply refer to it as the instantaneous frequency (IF) (Boashash, 1992). Note that for the spectra at the bottom of Figure 1, the pure harmonic frequencies of a trumpet cause the wrapped phase spectra to oscillate at different rates while the unwrapped phase smoothly diverges and the IF spectra forms solid bands where the harmonic frequencies are present.
在图 1 的中间，我们绘制了另一种方法，在相位语音编码器的启发下松散地生成相干波形。纯音产生一个进动的相位。在相位不连续时，通过增加 2π 来校正相位角，使进动的相位线性增长。然后，我们观察到，校正相位对于时间的导数保持不变，等于帧步长和信号周期之间的角度差。这通常被称为瞬时角频率，是真实信号振荡的时变度量。稍加滥用术语，我们只需将其称为瞬时频率（IF）。注意，对于图 1 底部的光谱，小号的纯谐波频率会导致包裹相位的频谱以不同的速率振荡，而未包裹相平滑地分离，IF 频谱形成存在谐波频率的实心带。

1.3 CONTRIBUTIONS
1.3 贡献

In this paper, we investigate the interplay of architecture and representation in synthesizing coherent audio with GANs. Our key findings include:
Generating log-magnitude spectrograms and phases directly with GANs can produce more coherent waveforms than directly generating waveforms with strided convolutions.
在本文中，我们研究了网络结构和表示方法在 GAN 合成相干音频时的相互作用。我们的主要发现包括：使用 GAN 生成对数级声谱图和相位，比直接用跨步卷积生成的波形更相干。
Estimating IF spectra leads to more coherent audio still than estimating phase.
It is important to keep harmonics from overlapping. Both increasing the STFT frame size and switching to mel frequency scale improve performance by creating more separation between the lower harmonic frequencies. Harmonic frequencies are multiples of the fundamental, so low pitches have tightly-spaced harmonics, which can cause blurring and overlap.
评估 IF（瞬时频率）光谱相比与评估相位，会产生更连贯的音频。
保持谐波不重叠非常重要。在更低的谐波频率之间的更多分离，比如增加 STFT 的帧大小、切换到 mel 频率刻度可提高性能。谐波频率是基础频率的倍数，因此低音调具有紧密间隔的谐波，这可能导致模糊和重叠。
On the NSynth dataset, GANs can outperform a strong WaveNet baseline in automatic and human evaluations, and generate examples ~54,000 times faster.
Global conditioning on latent and pitch vectors allow GANs to generate perceptually smooth interpolation in timbre, and consistent timbral identity across pitch.

在 NSynth 数据集中，GAN 在自动和人工评估中优于 WaveNet 基线，而且生成音频的速度为后者的 54, 000 倍。
对潜在和音高向量进行全局调节，使 GAN 能够在音色中生成感知平滑插值以及跨音高的一致音色标识。

2 EXPERIMENTAL DETAILS
2 实验细节

2.1 DATASET
2.1 数据集
We focus our study on the NSynth dataset, which contains 300,000 musical notes from 1,000 different instruments aligned and recorded in isolation. NSynth is a difficult dataset composed of highly diverse timbres and pitches, but it is also highly structured with labels for pitch, velocity, instrument, and acoustic qualities (Liu et al., 2015; Engel et al., 2017). Each sample is four seconds long, and sampled at 16kHz, giving 64,000 dimensions.
我们在 NSynth 数据集上工作，该数据集包含来自 1000 种独自对齐和录制的不同乐器演奏的 300，000 个音符。NSynth 是一个困难的数据集，由高度多样化的音色和音高组成，但它也高度结构化，具有音高、音量、乐器和声学质量的标签。每个样本有四秒长，以 16kHz 采样，共 64，000 个维度。
As we wanted to included human evaluations on audio quality, we restricted ourselves to training on the subset of acoustic instruments and fundamental pitches ranging from MIDI 24-84 (~32-1000Hz), as those timbres are most likely to sound natural to an average listener. This left us with 70,379 examples from instruments that are mostly strings, brass, woodwinds, and mallets. We created a new test/train 80/20 split from shuffled data, as the original split was divided along instrument type, which isn't desirable for this task.
由于我们希望包括人类对音频质量的评价，我们限制声学乐器和基本音高范围从 MIDI 24-84（~32-1000Hz），因为这些音色最有可能让人听起来感觉自然。这给我们留下了 70，379 个样本，乐器大多是弦乐，铜管乐，木管乐和打击乐。我们随机从数据中按 80/20 拆分创建了一个新的测试/训练集，因为原始的拆分是按仪器类型划分，不太适合我们的任务。

2.2 ARCHITECTURE AND REPRESENTATIONS
2.2 结构和表示

Taking inspiration from successes in image generation, we adapt the progressive training methods of Karras et al. (2018a) to instead generate audio spectra 3. While e search over a variety of hyper- parameter configurations and learning rates, we direct readers to the original paper for an in-depth analysis (Karras et al., 2018a), and the appendix for complete details.
从图像生成的成功中汲取灵感，我们调整 Karras 等人的渐进式训练方法，以生成音频声谱。在搜索各种超参数配置和学习速率时，我们建议读者到原始论文以及

附录进行深入分析，了解完整的详细信息。
Briefly, the model samples a random vector z from a spherical Gaussian, and runs it through a stack of transposed convolutions to upsample and generate output data x = G(z), which is fed into a discriminator network of downsampling convolutions (whose architecture mirrors the generator's) to estimate a divergence measure between the real and generated distributions (Arjovsky et al., 2017). As in Karras et al. (2018a), we use a gradient penalty (Gulrajani et al., 2017) to promote Lipschitz continuity, and pixel normalization at each layer. We also try training both progressive and non- progressive variants, and see comparable quality in both. While it is not essential for success, we do see slightly better convergence time and sample diversity for progressive training, so for the remainder of the paper, all models are compared with progressive training.
简而言之，该模型从球形高斯分布中采样随机向量 z，并通过一系列转置卷积操作，以向上采样并生成输出数据 x=G(z)，该数据被喂入下采样卷积（其网络结构和生成器对称）的鉴别器网络，以评估真实分布和生成的分布之间的发散程度。与 Karras 等人一样，我们使用梯度惩罚(gradient penalty)来促进 Lipschitz 连续性，并且将每个层的像素进行标准化。我们还尝试训练渐进式和非渐进式模型，并看到两者的质量相当。虽然不是特别的优势，但我们确实看到渐进式训练的收敛时间和样本多样性稍好一些，因此，在本文的其余部分中，所有模型都与渐进式训练进行比较。
Unlike Progressive GAN, our method involves conditioning on an additional source of information. Specifically, we append a one-hot representation of musical pitch to the latent vector, with the musically-desirable goal of achieving independent control of pitch and timbre. To encourage the generator to use the pitch information, we also add an auxiliary classification (Odena et al., 2017) loss to the discriminator that tries to predict the pitch label.
与渐进式 GAN 不同，我们的方法涉及到了对另一个信息源条件。具体来说，我们附加了一个音乐音调的 one-hot 向量，为了这个理想的目标，即实现对音调和音色的独立控制。为了鼓励生成器使用音调信息，我们还对鉴别器添加了一个辅助分类损失，用于预测音调标签。
For spectral representations, we compute STFT magnitudes and phase angles using TensorFlow's built-in implementation. We use an STFT with 256 stride and 1024 frame size, resulting in 75% frame overlap and 513 frequency bins. We trim the Nyquist frequency and pad in time to get an "image" of size (256, 512, 2). The two channel dimension correspond to magnitude and phase. We take the log of the magnitude to better constrain the range and then scale the magnitudes to be between -1 and 1 to match the tanh output nonlinearity of the generator network. The phase angle is also scaled to between -1 and 1 and we refer to these variants as "phase" models. We optionally unwrap the phase angle and take the finite difference as in Figure 1; we call the resulting models "instantaneous frequency" ("IF") models.
对于声谱表示，我们使用 TensorFlow 的内置实现计算 STFT 的幅度和相位角度。我们使用的 STFT 步长为 256，帧大小为 1024，这会导致 75%的帧重叠和 513 个频段。我们修剪 Nyquist 频率和补零操作得到一个大小为（256，512，2）的"图像"，数量为 2 通道对应于幅度和相位。 我们采用对数量级以更好地约束范围，然后缩放幅度在-1 和 1 之间，以匹配生成网络的 tanh 非线性输出。相位角也缩

放到-1 和 1 之间，我们将这些变体称为"相位"模型。我们可以选择解包相角，并采用如图 1 所示的有限差值;我们将生成的模型称为"瞬时频率"（"IF"）模型。
We also find performance is sensitive to having sufficient frequency resolution at the lower frequency range. Maintaining 75% overlap we are able to double the STFT frame size and stride, resulting in spectral images with size (128, 1024, 2), which we refer to as high frequency resolution, “+ H”, variants. Lastly, to provide even more separation of lower frequencies we transform both the log magnitudes and instantaneous frequencies to a mel frequency scale without dimensional compression (1024 bins), which we refer to as “IF-Mel” variants. To convert back to linear STFTs we just use the approximate inverse linear transformation, which, perhaps surprisingly does not harm audio quality significantly.
我们还发现模型的性能对在较低频率范围内是否具有足够的频率分辨率很敏感。保持 75%的重叠，STFT 帧大小和步幅翻倍，从而产生大小为(128,1024,2)的声谱图像，我们称之为高频分辨率，"+H"变体。最后，为了使更低的频率有更多分离，我们将对数级数和瞬时频率转换为没有尺寸压缩（1024 个分箱）的梅尔频率刻度，称之为"IF-Mel"变体。要转换回线性 STFT，只需要近似的逆线性变换，而且这并不会显著损害音频质量。
It is important for us to compare against strong baselines, so we adapt WaveGAN (Donahue et al., 2019), the current state of the art in waveform generation with GANs, to accept pitch conditioning and retrain it on our subset of the NSynth dataset. We also independently train our own waveform generating GANs off the progressive codebase and our best models achieve similar performance to WaveGAN without progressive training, so we opt to primarily show numbers from WaveGAN instead (see appendix Table 2 for more details).
与强基线进行比较非常重要，因此我们调整 WaveGAN，即具有 GAN 结构的波形生成网络，接受音调调节并在我们从 NSynth 数据集划出的子集上重新训练它。我们还独立训练自己的波形生成 GAN，我们最好的模型在没有渐进式训练的前提下达到和 WaveGAN 类似的性能，所以我们选择了展示 WaveGAN 的数据（请参阅附录表 2 有关更多的详细信息）。
Beyond GANs, WaveNet (van den Oord et al., 2016) is currently the state of the art in generative modeling of audio. Prior work on the NSynth dataset used an WaveNet autoencoder to interpolate between sounds (Engel et al., 2017), but is not a generative model as it requires conditioning on the original audio. Thus, we create strong WaveNet baselines by adapting the architecture to accept the same one-hot pitch conditioning signal as the GANs. We train variants using both a categorical 8-bit mu law and 16-bit mixture of logistics for the output distributions, but find that the 8-bit model is more stable and outperforms the 16-bit model (see appendix Table 2 for more details).
除了 GAN，WaveNet 目前是音频生成领域最先进的。NSynth 数据集的先前工作使用 WaveNet 自动编码器在声音之间插值，但不是生成模型，因为它需要对原始音频进行调节。因此，我们通过调整体系结构来接收与 GAN 相同的 one-hot 音高调理信号，从而创建更强的 WaveNet 基线。我们使用 categorical 8-bit mu law 和 16-bit mixture of logistics 来训练输出分布的变体，但发现 8-bit 模型更稳定且优于 16-bit 模型（有关详细信息，请参阅附录表 2）。

3 METRICS

3 评价指标

Evaluating generative models is itself a difficult problem: because our goals (perceptually-realistic audio generation) are hard to formalize, the most common evaluation metrics tend to be heuristic and have “blind spots” (Theis et al., 2016). To mitigate this, we evaluate all of our models against a diverse set of metrics, each of which captures a distinct aspect of model performance. Our evaluation metrics are as follows:

评估生成模型本身就是一个难题：因为我们的目标（感知真实的音频生成）很难正式化，最常见的评估指标往往是启发式的，并且有"盲点"。为了减轻这种情况，我们根据一组不同的指标评估所有模型，每个指标都捕获了模型性能的不同方面。我们的评估指标如下：

Human Evaluation We use human evaluators as our gold standard of audio quality because it is notoriously hard to measure in an automated manner. In the end, we are interested in training networks to synthesize coherent waveforms, specifically because human perception is extremely sensitive to phase irregularities and these irregularities are disruptive to a listener. We used Amazon Mechanical Turk to perform a comparison test on examples from all models presented in Table 1 (this includes the hold-out dataset). The participants were presented with two 4s examples corresponding to the same pitch. On a five-level Likert scale, the participants evaluate the statement Sample A has better audio quality / has fewer audio distortions than Sample B”. For the study, we collected 3600 ratings and each model is involved in 800 comparisons.

人类评估 我们使用人工评估作为音频质量的黄金标准，因为众所周知，很难以自动化的方式进行测量。最后，我们感兴趣的是训练网络来合成相干波形，特别是因为人类感知对相位不规则非常敏感，这些不规则对听者具有破坏性。我们使用 Amazon Mechanical Turk 对表 1 中介绍的所有模型的示例（包括保留数据集）执行比较测试。向参与者提供了两个与同一音高对应的 4s 示例。在五级类似表上，参与者评估样本 A 的具有比样本 B 更好的音频质量。在这项研究中，我们收集了 3600 个评分，每个模型都参与了 800 个比较。

Number of Statistically-Different Bins (NDB) We adopt the metric proposed by Richardson & Weiss (2018) to measure the diversity of generated examples: the training examples are clustered into k = 50 Voronoi cells by k-means in log-spectrogram space, the generated examples are also mapped into the same space and are assigned to the nearest cell. NDB is reported as the number of cells where the number of training examples is statistically significantly different from the number of generated examples by a two-sample Binomial test.

统计差异的 Bins 数量（NDB）我们采用 Richardson & Weiss 提出的指标来测量生成示例的多样性：在对数声谱图空间中，通过 k-means 将训练示例聚集到 k = 50 个 Voronoi 单元中，生成的示例也映射到同一空间并分配给最近的单元格。NDB 即为满足下面条件的单元数：单元中训练示例的数量在统计上与生成的示例数有显著差异。

Inception Score (IS) (Salimans et al., 2016) propose a metric for evaluating GANs which has become a de-facto standard in GAN literature (Gulrajani et al., 2017; Miyato et al., 2018; Karras et al., 2018a). Generated examples are run through a pretrained

Inception classifier and the Inception Score is defined as the mean KL divergence between the image- conditional output class probabilities and the marginal distribution of the same. IS penalizes models whose examples aren't each easily classified into a single class, as well as models whose examples collectively belong to only a few of the possible classes. Though we still call our metric "IS" for consistency, we replace the Inception features with features from a pitch classifier trained on spectrograms of our acoustic NSynth dataset.

初始分数（IS） Salimans 等人提出了一个评估 GAN 的指标，该指标已成为 GAN 文献中事实上的标准。生成的示例通过预训练的"初始"分类器运行，初始分数定义为图像-条件输出类概率和相同的边际分布之间的平均 KL 散度。IS 会惩罚其示例并非容易分类到单个类的模型，以及其示例都只属于少数可能类的模型。尽管我们仍将指标称为"IS"以表示一致性，但我们用经过声学 NSynth 数据集频谱训练的音高分类器的功能替换了初始特征。

Pitch Accuracy (PA) and Pitch Entropy (PE) Because the Inception Score can conflate models which don't produce distinct pitches and models which produce only a few pitches, we also separately measure the accuracy of the same pretrained pitch classifier on generated examples (PA) and the entropy of its output distribution (PE).

音高精度（PA）和音高熵（PE） 由于初始分数可以组合产生单一音调或少数种类音调的模型，我们还独立测量了生成示例在同一预训练音高分类器的准确性（PA）及其输出分布的熵（PE）。

Fre´chet Inception Distance (FID) (Heusel et al., 2017) propose a metric for evaluating GANs based on the 2-Wasserstein (or Fre´chet) distance between multivariate Gaussians fit to features extracted from a pretrained Inception classifier and show that this metric correlates with perceptual quality and diversity on synthetic distributions. As with Inception Score, we use pitch-classifier features instead of Inception features.

Fre_chet 初始距离（FID） Heusel 等人提出了一个指标，基于多变量高斯分布和从预先训练的 Inception 分类器中提取的特征的 2-Wasserstein(or Fre´chet)距离来评估 GAN，并表明该指标与生成分布的感知质量和多样性相关。与初始分数一样，我们使用音高分类器功能而不是初始特征。

4 RESULTS

4 结论

Table 1 presents a summary of our results on all model and representation variants. Our most discerning measure of audio quality, human evaluation, shows a clear trend, summarized in Figure 2. Quality decreases as output representations move from IF-Mel, IF, Phase, to Waveform. The highest quality model, IF-Mel, was judged comparably but slightly inferior to real data. The WaveNet base- line produces high-fidelity sounds, but occasionally breaks down into feedback and self oscillation, resulting in a score that is comparable to the IF GANs.

表 1 概述了我们关于所有模型和表示变体的结果。我们最关注的音频质量测量，人工评估，显示了一个明显的趋势，在图 2 中总结。随着输出表示从 IF-Mel、IF、相位到波形，质量会降低。最高质量模型 IF-Mel 被判断为相当，但略低于真实数据。WaveNet 基线产生高保真度声音，但偶尔会分解为反馈和自振荡，从而产

生与 IF GAN 相当的分数。

While there is no a priori reason that sample diversity should correlate with audio quality, we in- deed find that NDB follows the same trend as the human evaluation. Additionally, high frequency resolution improves the NDB score across models' types. The WaveNet baseline receives the worst NDB score. Even though the generative model assigns high likelihood to all the training data, the autoregressive sampling itself has a tendency gravitate to the same type of oscillation for each given pitch conditioning, leading to an extreme lack of diversity. Histograms of the sample distributions showing peaky distributions for the different models can be found in the appendix.

虽然没有先验的原因为什么样品多样性应该与音频质量相关，但我们确实发现了 NDB 遵循了和人类评估相同的趋势。此外，高频分辨率提高了不同模型的 NDB 分数。WaveNet 基线得到了最差的 NDB 分数。尽管生成模型为所有训练数据分配了很高的可能性，但自回归采样本身具有倾向性，即每个给定的音调都倾向于相同类型的振荡，从而导致极度缺乏多样性。在附录中可以找到显示不同模型样本分布的直方图。

Table 1: Metrics for different models. "+ H" stands for higher frequency resolution, and" Real Data" is drawn from the test set.

表 1：不同模型的指标。"+ H"代表更高的频率分辨率，"真实数据"是从测试集中提取的。

Examples	Human Eval (wins)	NDB	FID	IS	PA	PE
Real Data	549	2.2	13	47.1	98.2	0.22
IF-Mel + H	**485**	**29.3**	167	38.1	97.9	0.40
IF + H	308	36. 0	**104**	**41.6**	**98.3**	**0.32**
Phase + H	225	37.6	592	36.2	97.6	0.44
IF-Mel	479	37.0	600	29.6	94.1	0.63
IF	283	37.0	708	36.3	96.8	0.44
Phase	203	41.4	687	24.4	94.4	0.77
WaveNet	359	45.9	320	29.1	92.7	0.70
WaveGAN	216	43.0	461	13.7	82.7	1.40

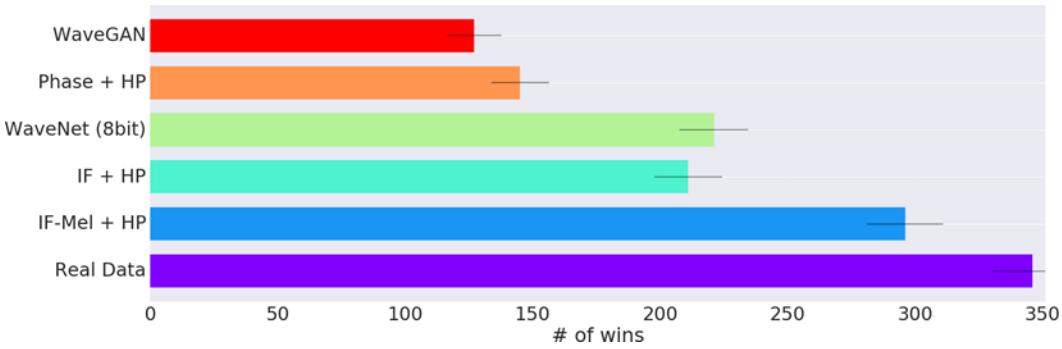

Figure 2: Number of wins on pair-wise comparison across different output representations and base- lines. Ablation comparing highest performing models of each type. Higher scores represent better perceptual quality to participants. The ranking observed here correlates well with the evaluation on quantitative metrics as in Table 1.

图 2：不同输出表示和基线上对比较的胜数。参与比较的是每种类型性能最高的模型。分数越高，参与者的感知质量就越高。此处观察到的排名与表 1 中对定量

指标的评估密切相关。

FID provides a similar story to the first two metrics with significantly lower scores for IF models with high frequency resolution. Comparatively, Mel scaling has much less of an effect on the FID then it does in the listener study. Phase models have high FID, even at high frequency resolution, reflecting their poor sample quality.
FID 与前两个指标相似，对于具有高频分辨率的 IF 模型，分数明显较低。相比之下，梅尔缩放相比于在听者研究中，对 FID 的影响要小得多。相位模型具有高 FID，即使在高频分辨率下，也反映出其样品质量差。

Many of the models do quite well on the classifier metrics of IS, Pitch Accuracy, and Pitch Entropy, because they have explicit conditioning telling them what pitches to generate. All of the high-resolution models actually generate examples classified with similar accuracy to the real data. As this accuracy and entropy can be a strong function of the distribution of generated examples, which most certainly does not match the training distribution due to mode collapse and other issues, there is little discriminative information to gain about sample quality from differences among such high scores. The metrics do provide a rough measure of which models are less reliably generating classifiable pitches, which seems to be the low frequency models to some extent and the baselines.
许多模型在 IS、音高精度和音高熵这些分类指标上表现相当出色，因为它们具有明确的条件，告诉它们要生成什么音调。所有高分辨率模型实际上都生成与实际数据类似的分类精度的示例。由于这种准确性和熵是生成示例分布的函数，这肯定与模型崩塌和其他问题导致的训练分布不匹配，因此，从如此高分数之间的差异中获得关于样本质量的区分信息很少。这些指标确实对哪些模型可以生成音高可分的样本提供了粗略的度量，结果是低频模型和基线生成的音频质量较低。

5 QUALITATIVE ANALYSIS
5 定性分析

While we do our best to visualize qualitative audio concepts, we highly recommend the reader to listen to the accompanying audio examples provided at
https://goo.gl/magenta/ gansynth-examples.
尽管我们尽力可视化定性音频概念，但我们强烈建议读者收听随附的音频示例，https://goo.gl/magenta/ gansynth-examples。

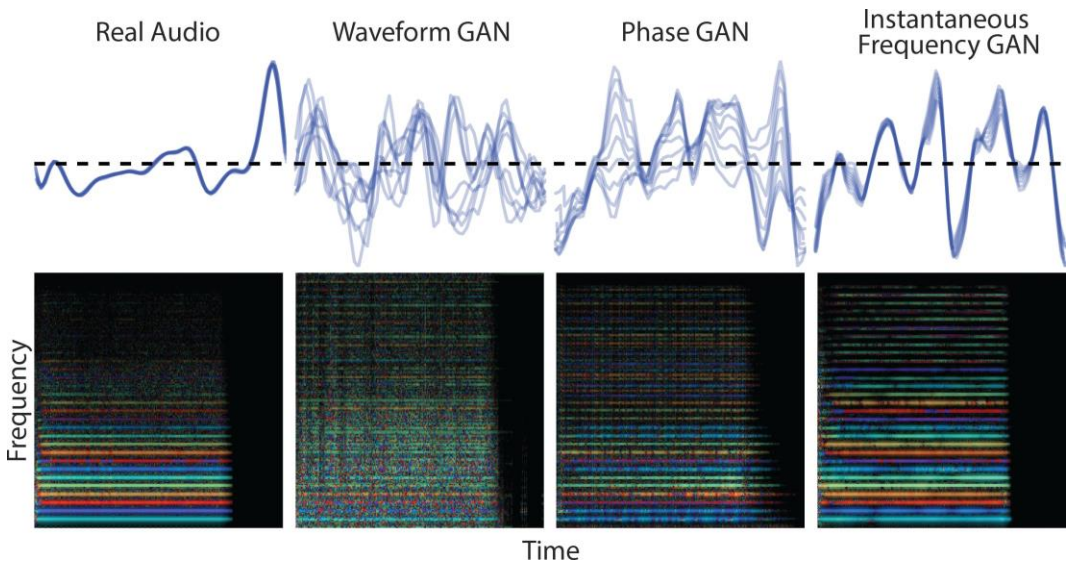

Figure 3: Phase coherence. Examples are selected to be roughly similar between the models for illustrative purposes. The top row shows the waveform modulo the fundamental periodicity of the note (MIDI C60), for 1028 examples taken in the middle of the note. Notice that the real data completely overlaps itself as the waveform is extremely periodic. The WaveGAN and PhaseGAN, however, have many phase irregularities, creating a blurry web of lines. The IFGAN is much more coherent, having only small variations from cycle-to-cycle. In the Rainbowgrams below, the real data and IF models have coherent waveforms that result in strong consistent colors for each harmonic, while the PhaseGAN has many speckles due to phase discontinuities, and the WaveGAN model is quite irregular.

图 3：相位一致性。为了说明目的，选择在模型之间大致相似的示例。最上一行显示了以音符的基本周期性为模的波形（MIDI C60），其中 1028 个示例位于音符的中间。注意，由于波形非常周期性，真实数据完全重叠。然而，WaveGAN 和 PhaseGAN 有许多相位不规则，形成了一个模糊的线网。IFGAN 更加协调，在循环到周期中只有很小的变化。在下面的彩虹图中，真实数据和 IF 模型具有相干波形，每个谐波产生强一致的颜色，PhaseGAN 由于相位不连续而有许多斑点，而 WaveGAN 模型非常不规则。

5.1 PHASE COHERENCE

5.1 相位一致性

Figure 3 visualizes the phase coherence of examples from different GAN variants. It is clear from the waveforms at the top, which are wrapped at the fundamental frequency, that the real data and IF models produce waveforms that are consistent from cycle-to-cycle. The PhaseGAN has some phase discontinuities, while the WaveGAN is quite irregular. Below we use Rainbowgrams (Engel et al., 2017) to depict the log magnitude of the frequencies as brightness and the IF as the color on a rainbow color map. This visualization helps to see clear phase coherence of the harmonics in the real data and IFGAN by the strong consistent colors. In contrast, the PhaseGAN discontinuities appear as speckled noise, and the WaveGAN appears largely incoherent.

图 3 展示了来自不同 GAN 变体的示例的相位一致性。从以基频包裹的顶部波形

可以清楚地看到，真实数据和 IF 模型产生的波形在各个周期之间都是一致的。PhaseGAN 具有某些相位不连续性，而 WaveGAN 非常不规则。下面我们使用 Rainbowgrams 在彩虹色图上将频率的对数幅度描述为亮度，将 IF 描述为颜色。这种可视化有助于通过强烈一致的颜色查看实际数据和 IFGAN 中谐波的相位相干性。相比之下，PhaseGAN 的不连续性表现为有斑点的噪声，而 WaveGAN 则出现很大程度的不连贯。

5.2 INTERPOLATION
5.2 插值

As discussed in the introduction, GANs also allow conditioning on the same latent vector the entire sequence, as opposed to only short subsequences for memory intensive autoregressive models like WaveNet. WaveNet autoencoders, such as ones in (Engel et al., 2017), learn local latent codes that control generation on the scale of milliseconds but have limited scope, and have a structure of their own that must be modelled and does not fit a compact prior. In Figure 4 we take a pretrained WaveNet autoencoder 5 and compare interpolating between examples in the raw waveform (top), the distributed latent code of a WaveNet autoencoder, and the global code of an IF-Mel GAN.

如引言中所述，GAN 还允许对整个序列在相同的潜在向量上进行条件处理，而对于像 WaveNet 这样的内存密集型自回归模型而言，仅允许使用较短的子序列。WaveNet 自动编码器，可以学习局部潜在编码，以毫秒为单位来控制生成，但是范围有限，并且具有自己的结构，必须对其进行建模而没有合适的先验。在图 4 中，我们采用了预训练的 WaveNet 自动编码器，比较原始波形中的示例（顶部），WaveNet 自动编码器的分布式潜在编码以及 IF-Mel GAN 的全局编码之间的内插。

Interpolating the waveform is perceptually equivalent to mixing between the amplitudes of two distinct sounds. WaveNet improves upon this for the two notes by mixing in the space of timbre, but the linear interpolation does not correspond to the complex prior on latents, and the intermediate sounds have a tendency to fall apart, oscillate and whistle, which are the natural failure modes for a WaveNet model. However, the GAN model has a spherical gaussian prior which is decoded to produce the entire sound, and spherical interpolation stays well-aligned with the prior. Thus, the perceptual change during interpolation is smooth and all intermediate latent vectors are decoded to produce sounds without additional artifacts. As a more musical example, in the audio examples, we interpolate the timbre between 15 random points in latent space while using the pitches from the prelude to Bach's Suite No. 1 in G major 6. As seen in appendix Figure 7, the timbre of the sounds morph smoothly between many instruments while the pitches consistently follow the composed piece.

内插波形在感觉上等效于两种不同声音的振幅之间的混合。WaveNet 通过在音色空间中进行混合对此进行了改进，但是线性插值并不对应于潜在音符上的复杂先验，并且中间的声音倾向于散开、振荡这是模型本身的缺陷。然而，GAN 模型具有球高斯先验分布，可以对其进行解码以产生整个声音，并且球面插值与先验保持良好对齐。因此，内插期间的感知变化是平滑的，并且所有的中间潜在向量都被解码以产生没有附加构件的声音。作为一个更具音乐性的示例，在音频示例

中，我们使用从巴赫 1 号套曲的前奏在 G 大调 6 的音高，在潜在空间中的 15 个随机点之间插值音色。如附录图 7 所示，声音在许多乐器之间平滑地变形，而音高始终跟随组成的乐段。

5.3 CONSISTENT TIMBRE ACROSS PITCH
5.3 音调间一致的音色

While timbre slightly varies for a natural instrument across register, on the whole it remains consistent, giving the instrument its unique character. In the audio examples 7, we fix the latent conditioning variable and generate examples by varying the pitch conditioning over five octaves. It's clear that the timbral identity of the GAN remains largely intact, creating a unique instrument identity for the given point in latent space. As seen in appendix Figure 7, the Bach prelude rendered with a single latent vector has a consistent harmonic structure across a range of pitches.

尽管音色在各种自然乐器中略有不同，但总体而言仍保持不变，从而赋予该乐器独特的特征。在音频示例 7 中，我们固定了潜在条件变量，并通过在五个八度音阶上改变音调条件来生成示例。很明显，GAN 的音色特征基本保持不变，从而为潜在空间中的给定点创建了唯一的乐器特征。如附录图 7 所示，用单个潜矢量绘制的巴赫前奏在音高范围内具有一致的谐波结构。

6 FAST GENERATION
6 快速生成

One of the advantages of GANs with upsampling convolutions over autoregressive models is that the both the training and generation can be processed in parallel for the entire audio sample. This is quite amenable to modern GPU hardware which is often I/O bound with iterative autoregressive algorithms. This can be seen when we synthesize a single four second audio sample on a TitanX GPU and the latency to completion drops from 1077.53 seconds for the WaveNet baseline to 20 milliseconds for the IF-Mel GAN making it around 53,880 times faster. Previous applications of WaveNet autoencoders trained on the NSynth dataset for music performance relied on prerendering all possible sounds for playback due to the long synthesis latency 8. This work opens up the intrigu- ing possibility for realtime neural network audio synthesis on device, allowing users to explore a much broader pallete of expressive sounds.

与自回归模型相比，具有向上采样卷积的 GAN 的优势之一是，可以针对整个音频样本并行处理训练和生成，这对于现代 GPU 硬件而言是相当合适的。当我们在 TitanX GPU 上合成一个四秒钟的音频样本时，可以看出这一点，完成时间从 WaveNet 基准的 1077.53 秒下降到 IF-Mel GAN 的 20 毫秒，使其速度提高了约 53,880 倍。在 NSynth 数据集上经过训练的 WaveNet 自动编码器在以前的音乐演奏应用中，由于较长的合成延迟而依赖于预渲染所有可能的声音进行播放。这项工作为设备上的实时神经网络音频合成提供了诱人的可能性，使用户能够探索更广泛的表达音色。

7 RELATED WORK
7 相关工作

Much of the work on deep generative models for audio tends to focus on speech synthesis (van den Oord et al., 2018; Sotelo et al., 2017; Wang et al., 2017). These datasets require handling variable length conditioning (phonemes/text) and audio, and often rely on recurrent and/or autoregressive models for variable length inputs and outputs. It would be interesting to compare adversarial audio synthesis to these methods, but we leave this to future work as adapting GANs to use variable-length conditioning or recurrent generators is a non-trivial extension of the current work.
关于音频的深度生成模型的很多工作都集中在语音合成上。这些数据集需要处理可变长度条件（音素/文本）和音频，并且经常依赖于递归或自回归模型来进行可变长度的输入和输出。将对抗性音频合成与这些方法进行比较会很有趣，但是我们将其留给以后的工作，因为使 GAN 适应使用可变长度条件或递归生成是当前工作的重要内容。
In comparison to speech, audio generation for music is relatively under-explored. van den Oord et al. (2016) and Mehri et al. (2017) propose autoregressive models and demonstrate their ability to synthesize musical instrument sounds, but these suffer from the aforementioned slow generation. Donahue et al. (2019) first applied GANs to audio generation with coherent results, but fell short of the audio fidelity of autoregressive likelihood models.
与语音相比，音乐的音频生成相对探索的更少。van den Oord 等和 Mehri 等人提出了自回归模型并证明了它们合成乐器声音的能力，但是它们受到生成速度缓慢造成的困扰。 Donahue 等首次将 GAN 应用于具有一致结果的音频生成，但未达到自回归似然模型的音频保真度。
Our work also builds on multiple recent advances in GAN literature. Gulrajani et al. (2017) propose a modification to the loss function of GANs and demonstrate improved training stability and archi- tectural robustness. Karras et al. (2018a) further introduce progressive training, in which successive layers of the generator and discriminator are learned in a curriculum, leading to improved generation quality given a limited training time. They also propose a number of architectural tricks to further improve quality, which we employ in our best models.
我们的工作还基于 GAN 系列的多项最新进展。Gulrajani 等提出了对 GAN 损失函数的修改，并证明了改进的训练的稳定性和架构鲁棒性。 Karras 等进一步引入了渐进式训练，在过程中训练了生成器和鉴别器的连续层，在有限的训练时间下可以提高生成器的质量。他们还提出了一些架构技巧，以进一步提高质量，我们在最佳模型中采用了这些技巧。
The NSynth dataset was first introduced as a “CelebA of audio” (Liu et al., 2015; Engel et al., 2017), and used WaveNet autoencoders to interpolate between timbres of musical instruments, but with very slow sampling speeds. Mor et al. (2019) expanded on this work by incoporating an adversarial domain confusion loss to achieve timbre transformations between a wide range of audio sources. Defossez et al. (2018) achieve significant sampling speedups (2,500x) over wavenet autoencoders by training a frame-based regression model to map from pitch and instrument labels to raw waveforms. They consider a unimodal likelihood regression loss in log spectrograms and back- propagate through the STFT, which yeilds good frequency estimation, but provides no incentive to learn phase coherency or handle multimodal distributions.

Their architecture also requires a large amount of channels, slowing down sample generation and training.
NSynth 数据集最初被称为“音频的 CelebA”，并使用 WaveNet 自动编码器在乐器的音色之间进行插值，但采样速度非常慢。 Mor 等通过合并对抗域的混淆损失来扩展此工作，以实现各种音频源头之间的音色转换。 Defossez 等通过训练基于帧的回归模型，从音高和乐器标签映射到原始波形，实现了 WaveNet 自动编码器的显著采样加速（2,500 倍）。他们考虑了对数频谱图中的单峰似然回归损失，并通过 STFT 反向传播，这可以产生良好的频率估计，但并不能激励学习相位相干性或处理多峰分布。他们的架构还需要大量通道，从而减慢了样本生成和训练的速度。

8 CONCLUSION
8 结论

By carefully controlling the audio representation used for generative modeling, we have demon- strated high-quality audio generation with GANs on the NSynth dataset, exceeding the fidelity of a strong WaveNet baseline while generating samples tens of thousands of times faster. While this is a major advance for audio generation with GANs, this study focused on a specific controlled dataset, and further work is needed to validate and expand it to a broader class of signals including speech and other types of natural sound. This work also opens up possible avenues for domain transfer and other exciting applications of adversarial losses to audio. Issues of mode collapse and diversity common to GANs exist for audio as well, and we leave it to further work to consider combining adversarial losses with encoders or more straightforward regression losses to better capture the full data distribution.
通过仔细控制用于生成模型的音频表示，我们在 NSynth 数据集上演示了使用 GAN 进行高质量音频生成，超过了强 WaveNet 基线的保真度，同时有数万倍的生成速度。 尽管这是使用 GAN 生成音频的一项重大进步，但本研究着眼于特定的受控数据集，还需要进一步的工作来验证并将其扩展到包括语音和其他类型自然声音在内的更广泛的信号类别。这项工作还为领域迁移和音频对抗性损失的其他激动人心的应用打开了可能的途径。音频也存在 GAN 常见的模式崩溃和多样性不足问题，我们将做进一步工作以考虑将对抗性损失与编码器或更直接的回归损失相结合，以更好地捕获完整的数据分布。

REFERENCES
参考文献

（略）

附录
A MEASURING DIVERSITY ACROSS GENERATED EXAMPLES
A 衡量生成示例的多样性

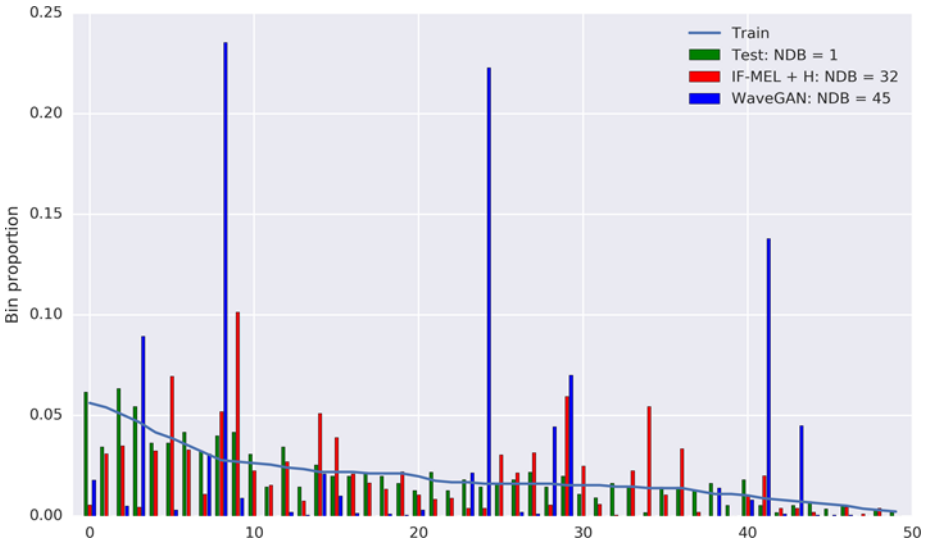

Figure 5: NDB bin proportions for the IF-Mel + H model and the WaveGAN baseline (evaluated with examples of pitch 60).

图 5：IF-Mel + H 模型和 WaveGAN 基线的 NDB bin 比例（使用音调 60 的示例评估）。

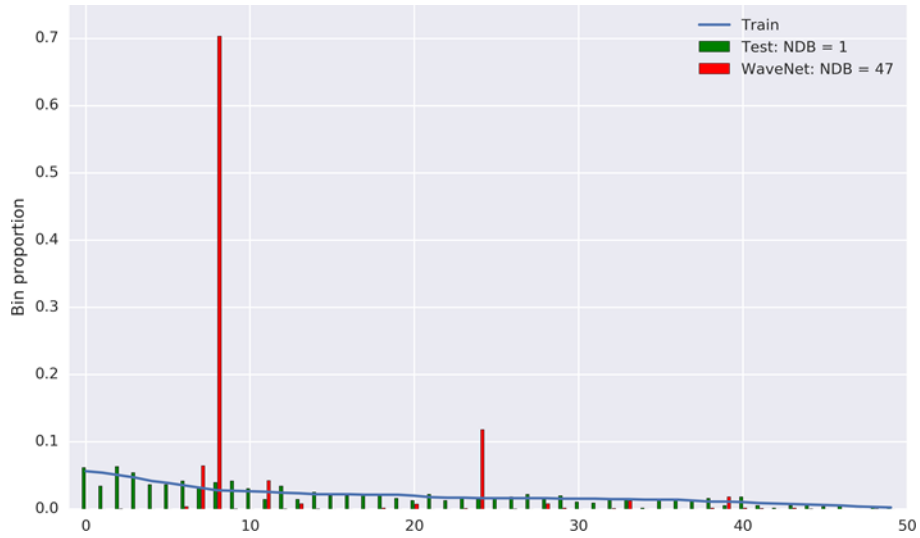

Figure 6: NDB bin proportions for the WaveNet baseline (evaluated with examples of pitch 60).

图 6：WaveNet 基线的 NDB bin 比例（使用音调 60 的示例评估）。

B TIMBRAL SIMILARITY ACROSS PITCH

B 跨音调的音色相似度

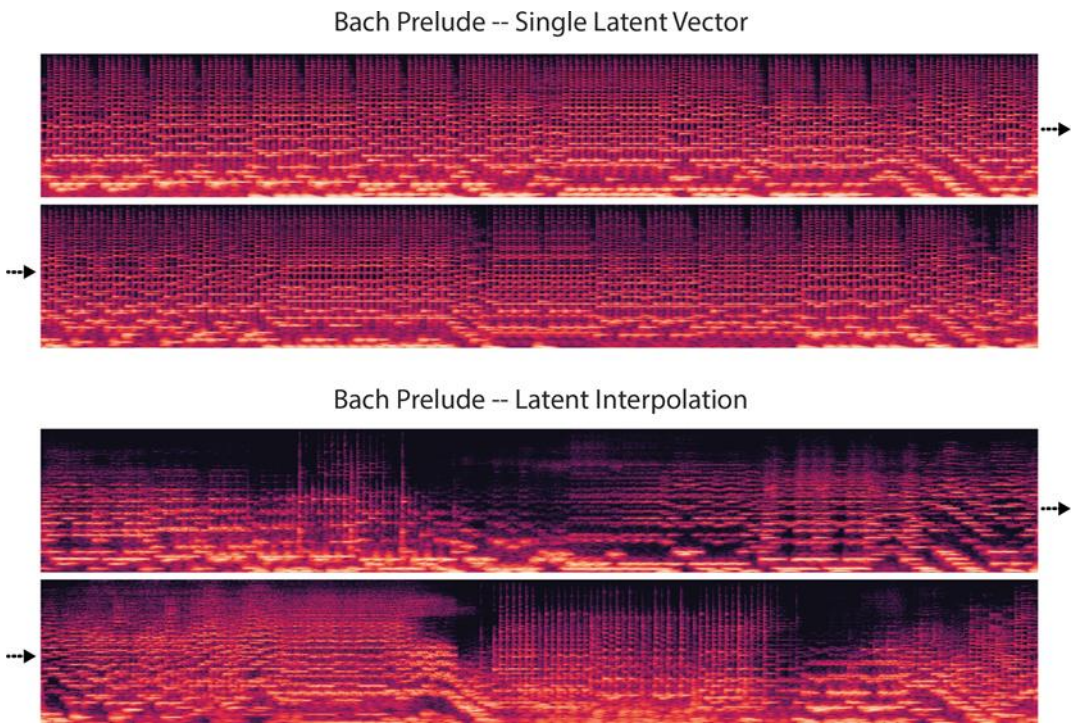

Figure 7: The first 20 seconds (10 seconds per a row) of the prelude to Bach's Suite No. 1 in G major 9, for pitches synthesized with a single latent vector (top), and with spherical interpolation in latent space (bottom). The timbre is constant for a single latent vector, shown by the consistency of the upper harmonic structure, while it varies dramatically as the latent vector changes. Listening examples are provided at https://goo.gl/magenta/gansynth-examples

图 7：G 大调 9 的巴赫第一序曲的前 20 秒（每行 10 秒），由单个潜在矢量合成的音高（上）以及在潜在空间中球形插值的合成（下）。对于单个潜矢量，音色是恒定的，如高次谐波结构的一致性所示，而随着潜矢量的变化，该音色会急剧变化。在 https://goo.gl/magenta/gansynth-examples 中提供了听力示例

C BASELINE MODEL COMPARISONS

C 基线模型比较

Table 2: Comparison of models generating waveforms directly. Our Waveform GAN baseline per- forms similar to the WaveGAN baseline, but the progressive training does not improve performance, so we only compare to the WaveGAN baseline for the paper. The 8-bit categorical WaveNet outperforms the 16-bit mixture of logistics, likely due to the decreased stability of the 16-bit model with only pitch conditioning, despite the increased fidelity.

表 2：直接生成波形的模型比较。我们的 Waveform GAN 基线的性能类似于 WaveGAN 基线，但是渐进式训练并不能提高性能，因此我们仅将本文与 WaveGAN 基线进行比较。8-bit categorical WaveNet 优于 16-bit mixture of logistics，这可能是由于尽管保真度提高了，但仅采用音调条件的 16 位模型的稳定性降低了。

Examples	NDB	FID	IS	PA	PE
WaveGAN	43.0	461	13.7	82.7	1.40
Waveform NoProg	48.2	447	14.8	96.3	1.61
Waveform Prog	45.0	375	2.5	56.7	3.59
WaveNet 8-bit	44.8	320	29.1	92.7	0.70
WaveNet 16-bit	45.9	656	9.5	64.6	1.71

D TRAINING DETAILS
D 训练细节

GAN architectures were directly adapted from an open source implementation in Tensorflow 10. Full details are given in Table 3, including adding a pitch classifier to the end of the discriminator as in AC-GAN. All models were trained with the ADAM optimizer (Kingma & Ba, 2014). We sweep over learning rates (2e-4, 4e-4, 8e-4) and weights of the auxiliary classifier loss (0.1, 1.0, 10), and find that for all variants (spectral representation, progressive/no progressive, frequency resolution) a learning rate of 8e-4 and classifier loss of 10 perform the best.
GAN 体系结构直接从 Tensorflow 中的开源代码实现改编而来。表 3 中给出了完整的详细信息，包括在 AC-GAN 中向标识符的末尾添加音高分类器。所有模型均使用 ADAM 优化器进行了训练。我们遍历学习率（2e-4、4e-4、8e-4）和辅助分类器损失的权重（0.1、1.0、10），发现对于所有变体（频谱表示，渐进/非渐进，频率分辨率），学习率 8e-4 和分类器损失 10 表现最佳。
As in the original progressive GAN paper, both networks use box upscaling/downscaling and the generators use pixel normalization,
与原始的渐进式 GAN 论文一样，两个网络都使用框式放大/缩小，而生成器使用像素归一化，

$$x = \frac{x_{nhwc}}{(\frac{1}{C}x^2)^{0.5}}$$

where n, h, w, and c refer to the batch, height, width, and channel dimensions respectively, x is the activations, and C is the total number of channels. The discriminator also appends the standard deviation of the minibatch activations as a scalar channel near the end of the convolutional stack as seen in Table 3.
其中 n，h，w 和 c 分别表示批次，高度，宽度和通道数，x 是 activations，C 是通道总数。鉴别器还将小批量 activations 的标准差附加为卷积层末端附近的标量通道，如表 3 所示。（？？？）
Since we find it helpful to use a Tanh output nonlinearity for the generator, we normalize real data before passing to the discriminator. We measure the maximum range over 100 examples and inde- pendently shift and scale the log-magnitudes and phases to [-0.8, 0.8] to allow for outliers and use more of the linear regime of the Tanh nonlinearity.
由于我们发现将 Tanh 非线性输出用于生成器是有帮助的，因此我们在传递给鉴别器之前将实数据标准化。我们测量了 100 个示例的最大范围，并独立地将对数幅度和相位移动并缩放至[-0.8，0.8]以允许出现异常值，并使用更多的 Tanh 非线

性机制。

We train each GAN variant for 4.5 days on a single V100 GPU, with a batch size of 8. For non- progressive models, this equates to training on 5M examples. For progressive models, we train on 1.6M examples per a stage (7 stages), 800k during alpha blending and 800k after blending. At the last stage we continue training until the 4.5 days completes. Because the earlier stages train faster, the progressive models train on ~11M examples.

我们在单个 V100 GPU 上对每个 GAN 变体进行了 4.5 天的训练，批处理大小为 8。对于非渐进模型，这相当于对 5M 示例进行训练。对于渐进模型，我们每个阶段（7 个阶段）训练 160 万个示例，alpha 混合期间训练 800k，混合之后训练 800k。在最后阶段，我们将继续训练直至 4.5 天结束。由于早期阶段的训练速度较快，因此渐进模型可以训练约 1100 万个示例。

For the WaveNet baseline, we also adapt the open source Tensorflow implementation 11. The decoder is composed of 30 layers of dilated convolution, each of 512 channels and receptive field of 3, and each with a 1x1 convolution skip connection to the output. The layers are divided into 3 stacks of 10, with dilation in each stack increasing from 20 to 29, and then repeating.

对于 WaveNet 基线，我们还改编了开源 Tensorflow 实现。该解码器由 30 层空洞卷积组成，每层 512 个通道，感受野为 3，并且通过 1*1 卷积和输出形成跳越连接。将这些层分为 3 块，每块 10 层，每个层的 dilation 从 20 增加到 29，然后重复。

We replace the audio encoder stack with a conditioning stack operating on a one-hot pitch conditioning signal distributed in time (3 seconds on, 1 second off). The conditioning stack is 5 layers of dilated convolution, increasing to 25, and then 3 layers of regular convolution, all with 512 channels. This conditioning signal is then passed through a 1x1 convolution for each layer of the decoder and added to the output of each layer, as in other implementations of WaveNet conditioning. For the 8-bit model we use mu- law encoding of the audio and a categorical loss, while for the 16-bit model we use a quantized mixture of 10 logistics (Salimans et al., 2017). WaveNets converged to 150k iterations in 2 days with 32 V100 GPUs trained with synchronous SGD with batch size 1 per GPU, for a total batch size of 32.

我们将音频编码器堆栈替换为操作在 one-hot 音调条件信号分布随时间的条件栈（3 秒钟打开，1 秒关闭）。条件堆栈是 5 层扩张卷积，增加到 25，然后是 3 层常规的卷积，全部具有 512 个通道。然后，与 WaveNet 条件的其他实现方式一样，此调理信号将通过解码器每一层的 1x1 卷积，并添加到每一层的输出中。对于 8-bit 模型，我们使用音频的 mu law 编码和分类损失，而对于 16 位模型，我们使用 a quantized mixture of 10 logistics。WaveNets 使用 32 个 V100 GPU 同步 SGD 训练，在 2 天内 150k 迭代收敛，每个 GPU 的批处理大小为 1，总批处理大小为 32。

Table 3: Model architecture for hi-frequency resolution. Low frequency resolution starts with a width of 4, and height of 8, but is otherwise the same. ”PN” stands for pixel norm, and ”LReLU” stands for leaky rectified linear unit, with a slope of 0.2. The latent vector Z has 256 dimensions and the pitch conditioning is a 61 dimensional one-hot vector.

表 3：高频分辨率的模型架构。低频分辨率的起始宽度为 4，高度为 8，其他方面相同。 “PN”代表像素范数，“LReLU”代表 leaky ReLU，斜率为 0.2。潜在向量 Z 有 256 维，音高条件为 61 维的 ont-hot 矢量。

Generator	Output Size	k_{Width}	k_{Height}	$k_{Filters}$	Nonlinearity
concat(Z, Pitch)	(1, 1, 317)	-	-	-	-
conv2d	(2, 16, 256)	2	16	256	PN(LReLU)
conv2d	(2, 16, 256)	3	3	256	PN(LReLU)
upsample 2x2	(4, 32, 256)	-	-	-	-
conv2d	(4, 32, 256)	3	3	256	PN(LReLU)
conv2d	(4, 32, 256)	3	3	256	PN(LReLU)
upsample 2x2	(8, 64, 256)	-	-	-	-
conv2d	(8, 64, 256)	3	3	256	PN(LReLU)
conv2d	(8, 64, 256)	3	3	256	PN(LReLU)
upsample 2x2	(16, 128, 256)	-	-	-	-
conv2d	(16, 128, 256)	3	3	256	PN(LReLU)
conv2d	(16, 128, 256)	3	3	256	PN(LReLU)
upsample 2x2	(32, 256, 256)	-	-	-	-
conv2d	(32, 256, 128)	3	3	128	PN(LReLU)
conv2d	(32, 256, 128)	3	3	128	PN(LReLU)
upsample 2x2	(64, 512, 128)	-	-	-	-
conv2d	(64, 512, 64)	3	3	64	PN(LReLU)
conv2d	(64, 512, 64)	3	3	64	PN(LReLU)
upsample 2x2	(128, 1024, 64)	-	-	-	-
conv2d	(128, 1024, 32)	3	3	32	PN(LReLU)
conv2d	(128, 1024, 32)	3	3	32	PN(LReLU)
generator output	(128, 1024, 2)	1	1	2	Tanh
Discriminator					
image	(128, 1024, 2)	-	-	-	-
conv2d	(128, 1024, 32)	1	1	32	-
conv2d	(128, 1024, 32)	3	3	32	LReLU
conv2d	(128, 1024, 32)	3	3	32	LReLU
downsample 2x2	(64, 512, 32)	-	-	-	-
conv2d	(64, 512, 64)	3	3	64	LReLU
conv2d	(64, 512, 64)	3	3	64	LReLU
downsample 2x2	(32, 256, 64)	-	-	-	-
conv2d	(32, 256, 128)	3	3	128	LReLU
conv2d	(32, 256, 128)	3	3	128	LReLU
downsample 2x2	(16, 128, 128)	-	-	-	-
conv2d	(16, 128, 256)	3	3	256	LReLU
conv2d	(16, 128, 256)	3	3	256	LReLU
downsample 2x2	(8, 64, 256)	-	-	-	-
conv2d	(8, 64, 256)	3	3	256	LReLU
conv2d	(8, 64, 256)	3	3	256	LReLU
downsample 2x2	(4, 32, 256)	-	-	-	-
conv2d	(4, 32, 256)	3	3	256	LReLU
conv2d	(4, 32, 256)	3	3	256	LReLU
downsample 2x2	(2, 16, 256)	-	-	-	-
concat(x, minibatch std.)	(2, 16, 257)	-	-	-	-
conv2d	(2, 16, 256)	3	3	256	LReLU
conv2d	(2, 16, 256)	3	3	256	LReLU
pitch classifier	(1, 1, 61)	-	-	61	Softmax
discriminator output	(1, 1, 1)	-	-	1	-