GANSYNTH: ADVERSARIAL NEURAL AUDIO SYNTHESIS

百度译文：对抗性神经音频合成

Google 译文：GANSYNTH：不可思议的神经音频合成

Jesse Engel, Kumar Krishna Agrawal, Shuo Chen, Ishaan Gulrajani, Chris Donahue, & Adam Roberts Google AI Mountain View, CA 94043, USA

百度译文：Jesse Engel、Kumar Krishna Agrawal、Shuo Chen、Ishaan Gulrajani、Chris Donahue 和 Adam Roberts Google Ai Mountain View，CA 94043，美国

Google 译文：Jesse Engel，Kumar Krishna Agrawal，Shuo Chen，Ishaan Gulrajani，Chris Donahue 和 Adam Roberts Google AI Mountain View，CA 94043，USA

ABSTRACT

Efficient audio synthesis is an inherently difficult machine learning task, as hu- man perception is sensitive to both global structure and fine-scale waveform co- herence. 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 paral- lel 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 investi- gations on the NSynth dataset, we demonstrate that GANs are able to outperform strong WaveNet baselines on automated and human evaluation metrics, and effi- ciently generate audio several orders of magnitude faster than their autoregressive counterparts.1

百度译文：有效的音频合成是一项本质上不同的机器学习任务，因为 Hu-man 感知对全局结构和线性波形一致性都很敏感。自回归模型，如 Wavenet、模型局部结构，但具有缓慢的迭代采样和缺乏全局潜在结构。相比之下，生成对抗网络（gans）具有全局潜在条件和有效的并行采样，但难以生成局部一致的音频波形。在此，我们证明，通过模拟对数量级和瞬时频率，在频谱域中具有良好的频率分辨率，gans 实际上可以生成高精度和局部相干的音频。通过对 nsynth 数据集的大量实证研究，我们证明，在自动化和人工评估指标上，gan 能够优于强大的 wavenet 基线，并且有效地生成音频，比它们的自回归对应项快几个数量级。1

Google 译文：有效的音频合成是一项固有的困难的机器学习任务，因为人类感知对全局结构和精细波形的相关性都很敏感。自回归模型，如 WaveNet，对局部结构进行建模，但迭代采样速度慢，缺乏全局潜在结构。相比之下，生成对抗网络（GAN）具有全局潜在调节和有效的并行采样，但很难生成局部相干的音频波形。在这里，我们通过在频谱域中以足够的频率分辨率对对数幅度和瞬时频率建模来证明 GAN 实际上可以产生高保真度和本地

相干音频。通过对 NSynth 数据集的广泛实证研究，我们证明了 GAN 能够在自动和人工评估指标上超越强大的 WaveNet 基线，并且能够比其自回归对应物更快地生成几个数量级的音频。

Introduction

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). 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.

百度译文：神经音频合成，训练生成模型以有效地生成具有高灵敏度和全局结构的音频，是一个具有挑战性的开放性问题，因为它需要在至少 5 个数量级（~0.1 ms 至~100 s）上建模时间尺度。几乎完全由自回归模型（如 Wavenet）开创了最新技术的巨大进步，这些模型通过关注可能的最高级（单个音频样本）并依靠全球结构的外部调节信号来解决规模问题（van den oord 等人，2016）。这是以缓慢采样速度为代价的，因为它们依赖于不充分的原始采样来生成波形，一次一个音频采样。由于他们的高质量，许多研究已经进入加速生成，但这些方法引入了显著的开销，例如培训中学生网络或编写高度定制的低水平内核（van den oord 等人，2018；paine 等人，2016）。此外，由于这些大型模型在一个固定的时间尺度上运行，因此它们的自动编码器变体仅限于由于内存限制而建模局部潜在结构（Engel 等人，2017）。在频谱的另一端，生成对抗网络（Gans）（Goodfellow 等人，2014）在生成高分辨率图像方面最近取得了巨大成功（Radford 等人，2016；Arjovsky

等人，2017；Gulrajani 等人，2017；Berthelot 等人，2017；Kodali 等人，2017；Karras 等人，2018a；Miyato 等人，2018）。典型的 gans 通过在一个潜在向量上调节一堆转置卷积来实现有效的并行采样和全局潜在控制，音频 gans 的潜力进一步扩大，因为对抗性成本为可能在音频中有类似物的图像解锁了有趣的域转换（isola 等人，2017；zhu 等人 al.，2017；Wolf 等人，2017；Jin 等人，2017）。然而，尝试调整图像 gan 架构以直接方式生成波形（Donahue 等人，2019 年）未能达到与图像对应物相同的感知水平。

Google 译文：神经音频合成，训练生成模型以有效地产生具有高维度和全局结构的音频，是一个具有挑战性的开放性问题，因为它需要建模至少五个数量级（~0.1ms 至~100s）的时间尺度。现有技术的巨大进步几乎完全由自回归模型开创，例如 WaveNet，它通过关注可能的嵌套规模（单个音频样本）来解决尺度问题，并依靠外部条件信号来实现全局结构（van den Oord 等，2016）。这是以慢采样速度为代价的，因为它们依靠无效的祖先采样来一次生成一个音频采样波形。由于它们的高质量，许多研究已经用于加速生成，但是这些方法引入了显着的开销，例如培训中学生网络或编写高度定制的低级内核（van den Oord 等，2018; Paine et al。，2016）。此外，由于这些大型模型在一个时间尺度上运行，因此它们的自动编码器变体仅限于由于存储器限制而仅对局部潜在结构进行建模（Engel 等，2017）。另一方面，Generative Adversarial Networks（GANs）（Goodfellow et al。，2014）最近在生成高分辨率图像方面取得了巨大成功（Radford 等，2016; Arjovsky 等，2017; Gulrajani 等。，2017; Berthelot 等，2017; Kodali 等，2017; Karras 等，2018a; Miyato 等，2018）。典型的 GAN 通过在潜在向量上调整一堆转置的卷积来实现高效的并行采样和全局潜在控制。音频 GAN 的潜力进一步扩展，因为对抗成本已经解锁了可能在音频中具有类似物的图像的有趣域转换（Isola） et al。，2017; Zhu et al。，2017; Wolf et al。，2017; Jin et al。，2017）。然而，尝试使图像 GAN 架构以直接的方式生成波形（Donahue 等，2019）未能达到与图像对应物相同的感知能力水平。

Colab Notebook: http://goo.gl/magenta/gansynth-demo, Audio Examples: http://goo.gl/magenta/gansynth-examples, Code: http://goo.gl/magenta/gansynth-code

百度译文：colab 笔记本：http://goo.gl/magenta/gansynth-demo，音频示例：http://goo.gl/magenta/gansynth-examples，代码：http://goo.gl/magenta/gansynth-code

Google 译文：Colab 笔记本：http：//goo.gl/magenta/gansynth-demo，音频示例：http：//goo.gl/magenta/gansynth-examples，代码：http：//goo.gl/magenta/gansynth-code

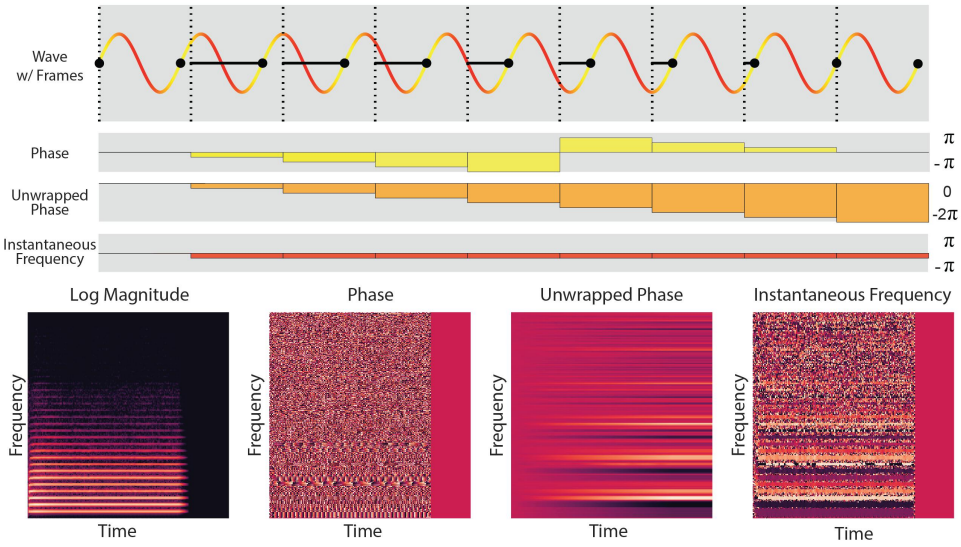

Figure 1: Frame-based estimation of audio waveforms. Much of sound is made up of locally-coherent waves with a local periodicity, pictured as the red-yellow sinusoid with black dots at the start of each cycle. Frame-based techniques, whether they be transposed convolutions or STFTs, have a given frame 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.

百度译文：图 1：基于帧的音频波形估计。许多声音都是由局部相干波和局部周期组成的，如每个周期开始时带有黑点的红黄色正弦曲线。基于帧的技术，无论是转置卷积还是 STFT，都有一个给定的帧大小和步幅，这里描述为与虚线的边界相等。由于音频和输出步幅的周期性不完全相同，这两个（相位，由实心黑线和黄色框表示）之间的对齐会及时进行。因此，转置卷积滤波器具有覆盖所有必要频率和所有可能相位对准以保持相位一致性的困难任务。对于 STFT，我们可以在 2π边界（橙色框）上展开相位，并取其导数得到瞬时径向频率（红色框），表示音频和帧频率之间的恒定关系。这些光谱显示为 NSynth 数据集的喇叭注释示例。

Google 译文：图 1：基于帧的音频波形估计。大部分声音由具有局部周期性的局部相干波组成，如每个周期开始时带有黑点的红黄色正弦波所示。基于帧的技术，无论它们是转置的卷积还是 STFT，都具有给定的帧大小和步幅，这里描述为与虚线处的边界相等。由于音频的周期性和输出步幅不完全相同，因此两者之间的对齐（相位，由实线黑线和黄色框表示）在时间上进行。因此，转置卷积滤波器具有覆盖所有必要频率和所有可能的相位对准以保持相位相干性的困难任务。对于 STFT，我们可以在 2π边界（橙色框）上展开相

位，并取其导数得到瞬时径向频率（红色框），表示音频和帧频之间的恒定关系。显示来自 NSynth 数据集的示例小号的光谱。

Generating Instrument Timbres

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 tex- ture and fine-scale features. Later models then built on that foundation to generalize to broader domains (Karras et al., 2018b; Brock et al., 2019). 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 instru- ments 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. 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.

百度译文：氮化镓研究人员通过对有限自由度的聚焦数据集的模型进行评估，逐步向不受约束的领域发展，从而在图像建模方面取得了快速进展。例如，流行的 Celeba 数据集（Liu et al.，2015）仅限于已居中和修剪的面，消除了姿势和姿势的变化，并为在生成真实的纹理和轮廓特征方面的质量改进（Radford et al.，2016；Karras et al.，2018a）提供了通用参考。后来的模型建立在该基础上，推广到更广泛的领域（KARAS 等人，2018B；Brink 等人，2019）。Nsynth 数据集（Engel 等人，2017）2 的引入具有相似的音频动机。NSynth 不包含所有类型的音频，而只包含来自不同音高、音色和音量的音乐机构的单个音符。与 Celeba 类似，所有数据都是对齐和裁剪的，以减少差异，并集中在与音色和音质相对应的音阶细节上。此外，每个注释还附带一组属性标签，以支持探索条件生成。最初的 NSynth 论文介绍了自回归波网自编码器和瓶颈谱图自编码器，但没有能力无条件地从先验数据中采样。后续工作探索了多种方法，包括基于帧的回归模型（Desossez 等人，2018 年）、反向散射网络（Andreux 和 Mallat，2018 年）、具有感知优先权的 Vaes（Esling 等人，2018 年）和域转移的对抗性规则化（Mor 等人，2019 年）。这项工作建立在这些努力的基础上，通过引入对抗性训练和探索非因果卷积生成的有效表示，如在甘氏综合症中发现的典型。

Google 译文：通过评估具有有限自由度的聚焦数据集上的模型，GAN 研究人员在图像建模方面取得了快速进展，并逐渐加强到受限较少的领域。例如，流行的 CelebA 数据集（Liu et al。，2015）仅限于已经居中和裁剪的面部，消除姿势和姿势的差异，并为定性改进提供共同参考（Radford 等，2016; Karras） et al。，2018a）生成逼真的纹理和细节特征。后来的模型建立在该基础之上，以推广到更广泛的领域（Karras 等，2018b; Brock 等，2019）。引入了 NSynth 数据集（Engel et al。，2017）2，具有类似的音频动机。NSynth 不是包含所有类型的音频，而是仅包含来自各种音高，音色和音量的音乐乐器的个别音符。与 CelebA 类似，所有数据都经过对齐和裁剪，以减少差异，并专注于细微的细节，音频对应于音色和细节。此外，每个音符还附有一系列属性标签，以便能够探索条件生成。最初的 NSynth 论文介绍了自回归 WaveNet 自动编码器和瓶颈谱图自动编码器，但没有能够无条件地从先前采样。后续工作探索了多种方法，包括基于框架的回归模型（Defossez 等，2018），逆散射网络（Andreux＆Mallat，2018），具有感知先验的 VAE（Esling 等，2018），以及对抗正则化。域转移（Mor et al。，2019）。这项工作建立在这些努力的基础上，通过引入对抗性训练和探索非因果卷积生成的有效表示，这是 GAN 中的典型特征。

Effective Audio Representations for GANs

Unlike images, most audio waveforms—such as speech and music—are highly periodic. Convolu- tional 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. This condition is assured as at any time there are typically many different frequencies in a given sig- nal. 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). 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 peri- odicity. 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.

百度译文：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。在这些数据上针对不同任务培训的卷积滤波器通常学习形成横跨人类听力范围的对数比例频率选择性滤波器组（Dieleman&Schrauwen，2014；Zhu 等人，2016）。人类的感知对周期波形的不连续和不规则也非常敏感，因此保持周期信号在短到中间时间尺度（1 ms-100 ms）内的规律性至关重要。图 1 显示，当帧的步幅不完全等于波形的周期性时，两个帧的对齐（相位）会随着时间推移而进行。这种情况是可以保证的，因为在任何时候，在一个给定信号中通常有许多不同的频率。这对合成网络来说是一个挑战，因为它必须学习所有合适的频率和相位组合，并以正确的组合激活它们，以产生一个连贯的波形。这种相位进动与短时间傅立叶变换（STFT）观测到的现象完全相同，后者由跨步滤波器组组成，就像卷积网络一样。相位进动也发生在滤波器组重叠的情况下（窗口或内核大小<跨距）。在图 1 的中间部分，我们描绘了另一种生成相干波形的方法，该方法受到相位声码器的松散启发（Dolson，1986）。纯音产生进动的相位。不包裹相位，只要它穿过一个相位不连续，就加上 2π，使进动相位线性增长。然后我们观察到展开相位对时间的导数保持不变，等于帧跨距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测量。稍微滥用一些术语，我们将其简单地称为瞬时频率（if）（Boashash，1992）。注意，对于图 1 底部的光谱，喇叭的纯谐波频率导致包裹的相位光谱以不同的速率振荡，而未包裹的相位平滑发散，并且如果光谱在存在谐波频率的地方形成固体带。

Google 译文：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。针对该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缩放频率选择滤波器组（Dieleman＆Schrauwen，2014; Zhu 等，2016）。人类感知对周期性波形中的不连续性和不规则性也非常敏感，因此在短到中间时间尺度（1ms-100ms）内保持周期性信号的规律性是至关重要的。图 1 显示，当帧的步幅不完全等于波形的周期性时，两个过程的对齐（相位）随时间推移。这种情况得到保证，因为在给定信号中通常存在许多不同的频率。这对于综合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，并以恰当的组合激活它们以产生相干波形。这种相位进动与短时傅里叶变换（STFT）观察到的现象完全相同，短时傅里叶变换由卷积网络组成的跨步滤波器组构成。相位进动也发生在滤波器重叠（窗口或内核大小<步幅）的情况下。在图 1 的中间，我们绘制了另一种方法来生成由相位声码器松散启发的相干波形（Dolson，1986）。纯音产生一个前进的阶段。通过在相位不连续处交叉时添加 2π来解开相位，导致进动阶段线性增长。然后我们观察到未展开相的时间相对于时间的导数保持不变，并且等于帧步幅和信号周期之间的角度差。这通常被称为瞬时角频率，并且是真实信号振荡的时变测量。稍微滥用术语，我们将其简称为瞬时频率（IF）（Boashash，1992）。注意，对于图 1 底部的光谱，喇叭的纯谐波频率导致包裹的相位谱以不同的速率振荡，而未包裹的相位平滑地发散，IF 频谱形成存在谐波频率的实心频带。

Contributions

In this paper, we investigate the interplay of architecture and representation in synthesizing coherent audio with GANs. Our key findings include:

百度译文：本文研究了用 GAN 合成相干音频时体系结构和表示的相互作用。我们的主要成果包括：

Google 译文：在本文中，我们研究了在与 GAN 合成相干音频时的架构和表示的相互作用。我们的主要发现包括：

• Generating log-magnitude spectrograms and phases directly with GANs can produce more

百度译文：•直接与 gans 生成对数幅谱图和相位可以产生更多

Google 译文：•直接用 GAN 生成对数幅度谱图和相位可以产生更多

coherent waveforms than directly generating waveforms with strided convolutions.

百度译文：相干波形比直接产生的跨步卷积波形。

Google 译文：相干波形比直接生成带有跨旋转的波形。

• 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.

百度译文：•估计频谱是否会导致比估计相位更连贯的音频。•重要的是防止谐波重叠。增加 STFT 帧大小和切换到 MEL 频率范围都可以通过在较低的谐波频率之间创建更多的分离来提高性能。谐波频率是乐趣达门塔尔的倍数，所以低间距有紧密间隔的谐波，这可能导致模糊和重叠。

Google 译文：•估计 IF 频谱导致比估计相位更连贯的音频。 •保持谐波不重叠很重要。增加 STFT 帧尺寸和切换到 mel 频率尺度都可以通过在较低谐波频率之间产生更多分离来改善性能。谐波频率是基波的倍数，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

百度译文：•在 nsynth 数据集上，gans 在自动和

Google 译文：•在 NSynth 数据集上，GAN 可以在自动和自动中优于强大的 WaveNet 基线

human evaluations, and generate examples ~54,000 times faster.

百度译文：人类评估，并更快生成 54000 倍的例子。

Google 译文：人工评估，并生成示例∽54,000 倍。

• Global conditioning on latent and pitch vectors allow GANs to generate perceptually

百度译文：•对潜在和音高向量的全局调节允许 gan 感知地生成

Google 译文：•潜在和俯仰矢量的全局调节允许 GAN 在感知上生成

smooth interpolation in timbre, and consistent timbral identity across pitch.

百度译文：音色平滑插值，音高一致。

Google 译文：音色中的平滑插值，以及音高之间一致的音色识别。

Experimental Details

Dataset

We focus our study on the NSynth dataset, which contains 300,000 musical notes from 1,000 differ- ent 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. As we wanted to included human evaluations on au- dio 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

百度译文：我们将研究重点放在 NSynth 数据集上，该数据集包含来自 1000 种不同乐器的 300000 种音符，这些乐器是单独排列和记录的。Nsynth 是一个由高度多样化的音色和音调组成的不同的数据集，但它也是高度结构化的，具有音调、速度、仪器和声学质量的标签（Liu 等人，2015；Engel 等人，2017）。每个样品长 4 秒，采样频率为 16kHz，尺寸为 64000。由于我们希望包括对音频质量的人类评估，我们将自己局限于对声学仪器和基本音高（从 MIDI 24-84（～32-1000Hz）的子集进行培训，因为这些音色最可能听起来自然。

Google 译文：我们将研究的重点放在 NSynth 数据集上，该数据集包含来自 1,000 个不同仪器的 300,000 个音符，这些乐器是单独对齐和记录的。 NSynth 是一个由高度多样化的音色和音高组成的难度数据集，但它也是高度结构化的音高，速度，乐器和声学质量的标签（Liu et al。，2015; Engel et al。，2017）。每个样品长 4 秒，以 16kHz 采样，得到 64,000 个尺寸。由于我们希望将人体评估纳入音频质量，我们仅限于使用 MIDI 24-84（∽32-1000Hz）的声学乐器和基本音高的训练，因为这些音色最有可能听起来很自然

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.

百度译文：对普通听众来说。这给我们留下了 70379 件乐器的例子，这些乐器大多是弦乐、铜管、木管乐器和木槌。我们从 Shuf fled 数据中创建了一个新的测试/列车 80/20 分割，因为最初的分割是沿着仪器类型进行的，这对于该任务来说是不可取的。

Google 译文：一般听众这给我们留下了 70,379 个例子，这些乐器主要是琴弦，黄铜，木管乐器和槌子。我们根据 shuf fl ed 数据创建了一个新的测试/训练 80/20 分割，因为原始分割是按照仪器类型划分的，这对于此任务是不可取的。

Architecture and Representations

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. 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. Unlike Progressive GAN, our method involves conditioning on an additional source of informa- tion. 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. 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. 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. 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). 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).

百度译文：从图像生成的成功经验中，我们采用了 Karras 等人的渐进式训练方法。（2018a）生成音频频谱 3。在搜索各种超参数配置和学习率的同时，我们将读者引向原始论文进行深入分析（Karras 等人，2018a），并在附录中了解完整的详细信息。Brie-fly，该模型从一个球面高斯中抽取一个随机向量 z，并将其穿过一组转置卷积进行上采样，并生成输出数据 x=g（z），该数据被送入一个下采样卷积的鉴别器网络（其结构反映了发生器的结构），以估计 real 和 ge 之间的散度度量。生成分布（Arjovsky 等人，2017）。如 Karras 等人所述。（2018a），我们使用梯度惩罚（Gulrajani 等人，2017）来促进 Lipschitz 连续性，并在每一层实现像素标准化。我们还尝试培训渐进式和非渐进式变体，并在两者中看到可比的质量。虽然这对成功并不重要，但我们发现渐进式训练的收敛时间和样本多样性稍好，因此在本文的其余部分，所有模型都与渐进式训练进行了比较。与渐进式的 gan 不同，我们的方法涉及对额外信息源的条件作用。具体地说，我们在潜在向量上附加了一个音乐音高的热表示，以实现音高和音色独立控制的音乐理想目标。为了鼓励发电机使用音高信息，我们还向试图预测音高标签的鉴别器添加了一个辅助分类（Odena 等人，2017）损失。对于光谱表示，我们使用 TensorFlow 的内置实现计算 stft 的大小和相位角。我们使用一个 256 步和 1024 帧大小的 STFT，导致 75%的帧重叠和 513 个频率箱。我们及时调整奈奎斯特频率和 PAD，以获得一个大小为（256、512、2）的"图像"。两个通道的尺寸对应于大小和相位。我们采用幅度对数来更好地约束范围，然后将幅度缩放到-1 和 1 之间，以匹配发电机网络的 tanh 输出非线性。相位角也被缩放到-1 和 1 之间，我们将这些变体称为"相位"模型。我们可以选择打开相位角，并采用如图 1 所示的有限差分；我们将所得模型称为"瞬时频率"（"if"）模型。我们还发现性能对在较低

的频率范围内获得足够的频率分辨率很敏感。保持 75%的重叠，我们能够将 STFT 帧大小和步幅增加一倍，从而产生大小为（128，1024，2）的光谱图像，我们称之为高频分辨率，“+h”，变体。最后，为了提供更低频率的更多分离，我们将对数量级和瞬时频率转换为无量纲压缩的 MEL 频率尺度（1024 箱），我们称之为“if-mel”变体。要转换回线性 stfts，我们只需要使用近似的反向线性转换，这可能令人惊讶地不会显著损害音频质量。比较强基线对我们来说很重要，因此我们采用 Wavegan（Donahue 等人，2019），这是 Gans 波形生成的最新技术，接受音调调节并在我们的 Nsynth 数据集子集上重传。我们还独立地从渐进式代码库中训练自己的波形生成 gan，并且我们的最佳模型在没有渐进式训练的情况下实现了与 wavegan 相似的性能，因此我们选择主要显示 wavegan 的数字（更多详情见附录表 2）。除了 Gans，Wavenet（van den Oord et al.，2016）目前是音频生成建模的最先进技术。之前对 nsynth 数据集的研究使用 WaveNet 自动编码器在声音之间进行插值（Engel 等人，2017），但不是生成模型，因为它需要对原始音频进行调节。因此，我们通过调整架构来接受与 gans 相同的一个热音高调节信号，从而创建强大的 wavenet 基线。我们使用分类 8 位 MU 定律和 16 位物流混合对输出分布进行培训，但发现 8 位模型更稳定，优于 16 位模型（更多详情见附录表 2）。

Google 译文：从图像生成的成功中汲取灵感，我们采用了 Karras 等人的渐进式训练方法。（2018a）改为生成音频谱 3.在搜索各种超参数配置和学习率时，我们引导读者阅读原始论文进行深入分析（Karras 等，2018a）和附录有关完整的细节。简而言之，该模型从球形高斯中采样随机向量 z，并通过一组转置的卷积运行它以进行上采样并生成输出数据 x = G（z），将其输入下采样卷积的鉴别器网络（其结构镜像）生成器）估计实际分布和生成分布之间的差异度量（Arjovsky 等，2017）。和 Karras 等人一样。（2018a），我们使用梯度惩罚（Gulrajani 等，2017）来促进 Lipschitz 连续性，以及每层的像素归一化。我们还尝试训练渐进和非渐进变体，并在两者中看到相当的质量。虽然它不是成功的必要条件，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，因此对于本文的其余部分，所有模型都与渐进式训练进行比较。与 Progressive GAN 不同，我们的方法涉及调整其他信息来源。具体而言，我们在潜在向量上添加了一个单一的音高表示，其音乐理想的目标是实现对音高和音色的独立控制。为了鼓励发电机使用音高信息，我们还向试图预测音高标签的鉴别器添加辅助分类（Odena 等，2017）损失。对于频谱表示，我们使用 TensorFlow 的内置实现计算 STFT 幅度和相位角。我们使用具有 256 步幅和 1024 帧尺寸的 STFT，导致 75％的帧重叠和 513 个频率区间。我们修剪奈奎斯特频率并及时填充以获得大小（256,512,2）的“图像”。两个通道尺寸对应于幅度和相位。我们采用幅度的对数来更好地约束范围，然后将幅度缩放到-1 和 1 之间，以匹配发电机网络的 tanh 输出非线性。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。我们可选择打开相位角并采用有限差分，如图 1 所示;我们称之为模型“瞬时频率”（“IF”）模型。我们还发现性能对在较低频率范围内具有足够的频率分辨率很敏感。保持 75％的重叠，我们能够将 STFT 帧尺寸和步幅加倍，从而产生尺寸为（128,1024,2）的光谱图像，我们将其称为高频分辨率“+ H”变体。最后，为了提供更低频率的分离，我们将对数幅度和瞬时频率转换为没有尺寸压缩（1024 个区间）的梅尔频率标度，我们将其称为“IF-Mel”变体。为了转换回线性 STFT，我们只使用近似逆线性变换，这可能令人惊讶地不会显着损害音频质量。

对于我们来说，与强基线进行比较非常重要，因此我们采用 WaveGAN（Donahue 等，2019），使用 GAN 进行波形生成的当前最新技术，接受音调调节并在 NSynth 数据集的子集上重新训练它。 。我们还独立训练我们自己的波形，从渐进式代码库中生成 GAN，我们的最佳模型在没有渐进式训练的情况下实现与 WaveGAN 类似的性能，因此我们选择主要显示来自 WaveGAN 的数字（有关详细信息，请参阅附录表 2）。除了 GAN 之外，WaveNet（van den Oord 等，2016）目前是音频生成建模的最先进技术。关于 NSynth 数据集的先前工作使用 WaveNet 自动编码器在声音之间进行插值（Engel 等，2017），但不是生成模型，因为它需要对原始音频进行调节。因此，我们通过调整架构来接受与 GAN 相同的单热调节信号来创建强大的 WaveNet 基线。我们使用分类 8 位μ律和 16 位物流混合来训练变量，用于输出分布，但发现 8 位模型更稳定并且优于 16 位模型（更多细节见附录表 2））。

Metrics

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:

百度译文：评估生成模型本身就是一个困难的问题：因为我们的目标（感知现实的音频生成）很难形式化，最常见的评估指标往往是启发式的，并且有“盲点”（Theis 等人，2016）。为了减轻这一点，我们根据一组不同的度量来评估我们的所有模型，每个度量都捕获了模型性能的不同方面。我们的评估指标如下：

Google 译文：评估生成模型本身就是一个难题：因为我们的目标（感知逼真的音频生成）难以形式化，最常见的评估指标往往是启发式的，并且具有“盲点”（Theis et al。，2016）。为了缓解这种情况，我们根据各种指标评估所有模型，每个指标都捕获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 be- cause 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 percep- tion 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 Lik- ert scale, the participants evaluate the statement ”Sample A has better audio quality / has less audio distortions than Sample B”. For the study, we collected 3600 ratings and each model is involved in 800 comparisons.

百度译文：•人性化评估我们将人性化评估作为我们音频质量的黄金标准，因为以自动化方式衡量是出了名的困难。最后，我们对训练网络合成相干波形很感兴趣，特别是因为人类感知对相位不规则非常敏感，这些不规则对听者来说是破坏性的。我们使用 AmazonMechanical Turk 对表 1 中给出的所有模型的示例进行了比较测试（这包括保持数据集）。向参加者展示了两个与同一音高相对应的 4s 例子。在五级水平上，参与者评估“样本 A 的音频质量更好/音频失真比样本 B 少”这一说法。在研究中，我们收集了 3600 个评级，每个模型都涉及 800 个比较。

Google 译文：•人体评估我们使用人工评估员作为音频质量的黄金标准，因为以自动方式测量是非常困难的。最后，我们感兴趣的是训练网络以合成相干波形，特别是因为人类感知对相位不规则性极为敏感，而这些不规则性对听者来说是破坏性的。我们使用 Amazon Mechanical Turk 对表 1 中列出的所有模型的示例进行了比较测试（这包括保留数据集）。向参与者呈现了对应于相同音高的两个 4s 示例。在五级标准尺度上，参与者评估声明“样本 A 具有更好的音频质量/具有比样本 B 更少的音频失真”。在这项研究中，我们收集了 3600 个评级，每个模型参与了 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.

百度译文：•统计上不同的箱数（ndb）我们采用 Richardson&Weiss（2018）提出的度量标准来衡量生成示例的多样性：训练示例通过对数谱图空间中的 k 平均数聚集成 k=50 个 Voronoi 单元，生成的示例也映射到相同的空间，并分配给最近的单元。.ndb 被报告为训练样本数量统计上显著不同于两样本二项式测试生成样本数量的细胞数量。

Google 译文：•统计上不同的区间数（NDB）我们采用 Richardson＆Weiss（2018）提出的度量来测量生成的例子的多样性：训练样例在对数谱图空间中通过 k 均值聚类成 k = 50 个 Voronoi 单元格，生成的示例也映射到相同的空间并分配给最近的单元格。 NDB 被报告为单元数，其中训练样本的数量在统计上与双样本二项式测试的生成示例的数量显着不同。

• Inception Score (IS)

百度译文：初始分数（IS）

Google 译文：•初始分数（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.

百度译文：（Salimans 等人，2016）提出了评估 gan 的指标，该指标已成为 gan 文献中的事实标准（Gulrajani 等人，2017；Miyato 等人，2018；Karras 等人，2018a）。生成的示例通过预先培训的初始分类进行运行，初始分数被定义为图像条件输出分类概率和边缘分布之间的平均 kl 差异。它惩罚的是不容易将每个示例分类为一个类的模型，以及其示例只属于几个可能的类的模型。尽管我们仍然称我们的度量标准为"IS"，以保持一致性，但我们将最初的特性替换为音高等级的特性，这些特性是根据我们的声学 NSynth 数据集的光谱图进行培训的。

Google 译文：（Salimans 等，2016）提出了评估 GAN 的指标，该指标已成为 GAN 文献中的事实标准（Gulrajani 等，2017; Miyato 等，2018; Karras 等，2018a）。生成的示例通过预训练的初始分类器运行，并且初始分数被定义为图像条件输出类概率与其的边际分布之间的平均 KL 偏差。 IS 会惩罚那些示例并非易于分类为单个类的模型，以及其示例共同仅属于少数可能类的模型。尽管我们仍然将度量标准称为"IS"以保持一致性，但我们将使用来自我们的声学 NSynth 数据集的频谱图训练的音高分类器的特征替换 Inception 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).

百度译文：•音高精度（PA）和音高熵（PE），因为初始分数可以影响不产生不同音高的模型和只产生几个音高的模型，所以我们还分别测量生成示例（PA）上相同预训练音高等级的精度及其输出分布的熵（PE）。

Google 译文：•俯仰精度（PA）和俯仰熵（PE）因为初始分数可以包含不产生不同节距的模型和仅产生几个节距的模型，我们还分别测量生成的示例中相同的预训练音高分类的精度（PA）和其输出分布（PE）的熵。

• Fr´echet Inception Distance (FID) (Heusel et al., 2017) propose a metric for evaluating GANs based on the 2-Wasserstein (or Fr´echet) 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.

百度译文：•Fr'Echet 起始距离（Fid）（Heusel 等人，2017 年）提出了一种基于从预培训起始分类中提取的特征的多变量 Gaussians 之间的 2-Wasserstein（或 Fr'Echet）距离的评估 Gans 的指标，并表明该指标与感知质量和合成分布的多样性有关。与初始分数一样，我们使用音高等级的特征而不是初始特征。

Google 译文：•Fr'echet 初始距离（FID）（Heusel 等，2017）提出了一个度量标准，用于根据多变量高斯算子之间的 2-Wasserstein（或 Fr'echet）距离来评估 GAN，以及从预训练的初始分类器中提取的特征并显示该指标与合成分布的感知质量和多样性相关。与 Inception Score 一样，我们使用 pitch-classi fi er 功能而不是 Inception 功能。

Results

Table 1 presents a summary of our results on all model and representation variants. Our most dis- cerning 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. 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

百度译文：表 1 总结了我们对所有模型和表示变量的结果。我们对音频质量最为关注的度量，即人的评估，显示出一个明显的趋势，如图 2 所示。当输出表示从 if-mel、if、phase 移动到 waveform 时，质量会降低。最高质量模型（如果是 MEL）的判断与实际数据相当，但略逊于实际数据。wavenet 基线产生高保真的声音，但偶尔会分解为反馈和自我振荡，从而得出与 if-gans 相当的分数。虽然没有先验的理由认为样本多样性与音频质量相关，但我们发现，NDB 与人类评估遵循相同的趋势。此外，高频分辨率提高了不同型号的 NDB 评分。wavenet 基线得到最差的 NDB 分数。尽管生成模型为所有训练数据赋予了很高的可能性，但是

Google 译文：表 1 总结了我们对所有模型和表示变体的结果。我们对音频质量的最有效测量，人工评估，显示了一个明显的趋势，如图 2 所示。随着输出表示从 IF-Mel，IF，Phase 到 Waveform，质量下降。最高质量的模型 IF-Mel 被评判为相对但略低于实际数据。 WaveNet 基线产生高音质声音，但偶尔会分解为反馈和自振荡，从而得分与 IF GAN 相当。虽然没有先验的理由认为样本多样性应与音频质量相关，但我们确实发现 NDB 遵循与人类评估相同的趋势。此外，高频率分辨率可提高不同模型类型的 NDB 分数。WaveNet 基线获得最差的 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，但是

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

百度译文：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。

Google 译文：表 1：不同模型的度量标准。 “+ H”代表更高的频率分辨率，“真实数据”代表测试集。

Human Eval

百度译文：人类时代

Google 译文：人类评估

Examples Real Data IF-Mel + H IF + H Phase + H IF-Mel IF Phase WaveNet WaveGAN

百度译文：如果 mel+h if+h phase+h if mel if phase wavenet wavegan，示例实数

Google 译文：实例数据实数数据 IF-Mel + H IF + H 相+ H IF-Mel IF 相位 WaveNet WaveGAN

(wins) 549 485 308 225 479 283 203 359 216

百度译文：（胜）549 485 308 225 479 283 203 359 216

Google 译文：（胜利）549 485 308 225 479 283 203 359 216

NDB FID 13 2.2 29.3 167 104 36. 0 592 37.6 600 37.0 708 37.0 41.4 687 320 45.9 43.0 461

百度译文：NDB FID 13 2.2 29.3 167 104 36.0 592 37.6 600 37.0 708 37.0 41.4 687 320 45.9 43.0 461

Google 译文：NDB IN 13 2.2 29.3 167 104 36 0 592 37.6 37.0 600 37.0 708 41.4 687 320 45.9 43.0 461

IS 47.1 38.1 41.6 36.2 29.6 36.3 24.4 29.1 13.7

百度译文：IS 47.1 38.1 41.6 36.2 29.6 36.3 24.4 29.1 13.7

Google 译文：IS 47.1 38.1 41.6 36.2 29.6 36.3 24.4 29.1 13.7

PA 98.2 97.9 98.3 97.6 94.1 96.8 94.4 92.7 82.7

百度译文：PA 98.2 97.9 98.3 97.6 94.1 96.8 94.4 92.7 82.7

Google 译文：PA 98.2 97.9 98.3 97.6 94.1 96.8 94.4 92.7 82.7

PE 0.22 0.40 0.32 0.44 0.63 0.44 0.77 0.70 1.40

百度译文：体育 0.22 0.40 0.32 0.44 0.63 0.44 0.77 0.70 1.40

Google 译文：PE 0.22 0.40 0.32 0.44 0.63 0.44 0.77 0.70 1.40

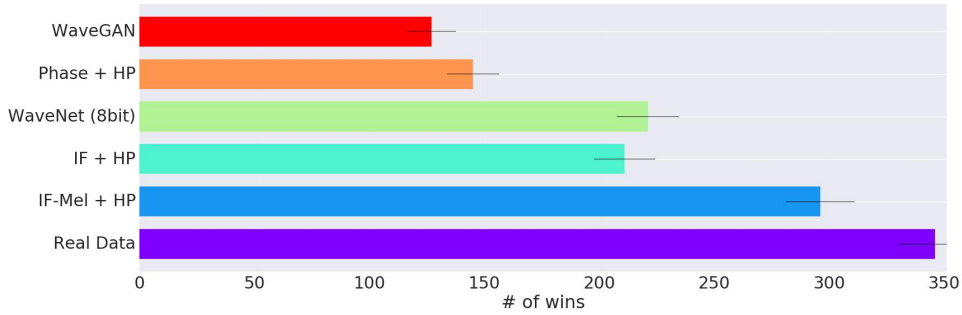

Figure 2: Number of wins on pair-wise comparison across different output representations and base- lines. Ablation comparing highest performing models of each type. Higher scores represent better perceptual quality to participants. The ranking observed here correlates well with the evaluation on quantitative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

百度译文：图 2：不同输出表示形式和基线之间的成对比较获胜数。消融比较每种类型的最高性能模型。分数越高，参与者的感知质量越好。这里观察到的排名与表 1 中的定量指标评估有很好的相关性。

Google 译文：图 2：不同输出表示和基线的成对比较的胜利数。消融比较每种类型的最高性能模型。分数越高表示参与者的感知质量越好。这里观察到的排名与定量指标的评估很好地相关，如表 1 所示。

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. FID provides a similar story to the first two metrics with significantly lower scores for 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. Many of the models do quite well on the classifier metrics of IS, Pitch Accuracy, and Pitch En- tropy, 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.

百度译文：自回归采样本身有一种趋势，即对于每个给定的音调条件，自回归采样本身都倾向于同一类型的振荡，从而导致极端缺乏多样性。显示不同模型峰值分布的样本分布柱状图见附录。fid 提供了与前两个指标类似的情况，对于具有高频率分辨率的 if 模型，分数明显较低。相比之下，MEL 缩放对 FID 的影响要小得多，而在侦听器研究中则是如此。相位模型具有高的 FID，即使在高频率分辨率下，也会重新反映出其糟糕的样品质量。许多模型在 IS、音高精度和音高各向异性等更高级的指标上做得很好，因为它们有明确的条件作用，告诉它们要生成什么音高。所有的高分辨率模型实际上都生成了与真实数据具有相似精度的分类示例。由于这种准确性和熵可以是生成样本分布的一个强大函数，由于模式崩溃和其他问题，很显然与训练分布不匹配，因此从这些高分之间的差异中获得样本质量的识别信息很少。度量标准确实提供了一个粗略的度量，其中哪些模型不太可靠地生成可分类的音高，在某种程度上，这些模型似乎是低频模型和基线。

Google 译文：对于每个给定的音调调节，自回归采样本身具有倾向于相同类型的振荡的趋势，导致极度缺乏多样性。显示不同模型的峰值分布的样本分布的直方图可以在附录中找到。 FID 为前两个指标提供了类似的故事，对于具有高频率分辨率的 IF 模型，得分显着较低。相比之下，梅尔缩放对 FID 的影响远小于听众研究中的影响。即使在高频率分辨率下，相位模型也具有高 FID，从而反映出其较差的样品质量。许多模型在 IS，Pitch Accuracy 和 Pitch Enropropy 的分类指标上做得很好，因为它们有明确的条件，告诉他们要产生什么样的音高。所有高分辨率模型实际上生成的示例分类与实际数据具有相似的精度。由于这种准确性和熵可能是生成的例子的分布的强大函数，由于模式崩溃和其他问题，这些例子肯定与训练分布不匹配，因此从这样的高分之间的差异中获得关于样本质量的判断信息很少。 。这些指标确实可以粗略地衡量哪些模型不太可靠地生成可分类的音高，这似乎是某种程度上的低频模型和基线。

Qualitative Analysis

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.

百度译文：虽然我们尽最大努力将定性音频概念可视化，但我们强烈建议读者仔细阅读 https://goo.gl/magenta/gansynth 示例中随附的音频示例。

Google 译文：虽然我们尽最大努力可视化定性音频概念，但我们强烈建议读者聆听 https://goo.gl/magenta/gansynth- examples 中提供的相应音频示例。

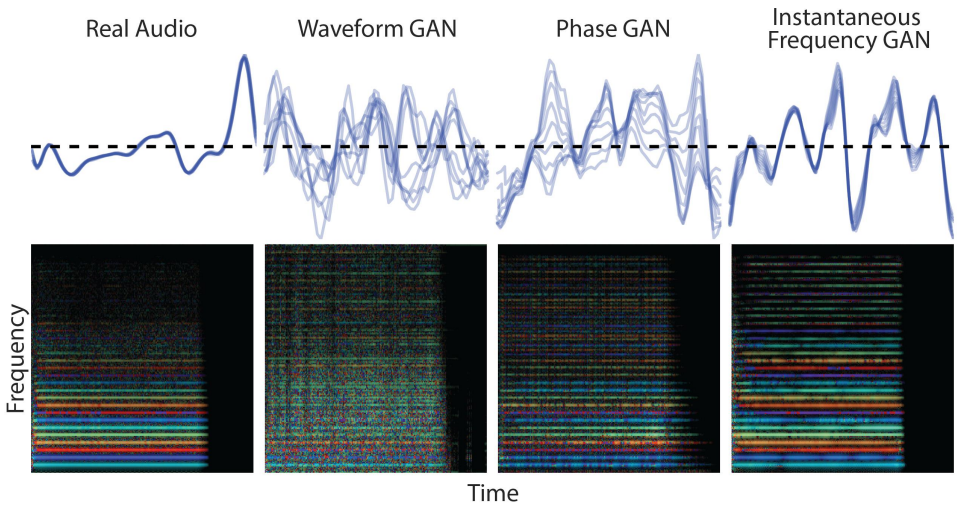

Figure 3: Phase coherence. Examples are selected to be roughly similar between the models for illustrative purposes. The top row shows the waveform modulo the fundamental periodicity of the note (MIDI C60), for 1028 examples taken in the middle of the note. Notice that the real data completely overlaps 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 har- monic, while the PhaseGAN has many speckles due to phase discontinuities, and the WaveGAN model is quite irregular.

百度译文：图 3：相位相干性。为了便于说明，选择的示例在模型之间大致相似。最上面一行显示了音符（MIDI C60）基本周期的波形模件，在音符中间有 1028 个例子。请注意，实际数据完全重叠，因为波形是非常周期的。然而，Wavegan 和 Phasegan 有许多相位不规则，形成了一个模糊的线网。ifgan 的一致性更高，从一个周期到另一个周期只有很小的变化。在下面的彩虹图中，真实的数据和中频模型都有相干的波形，这些波形会使每个 har monic 产生强烈的一致的颜色，而相位发生器由于相位不连续而有许多斑点，并且 Wavegan 模型是非常不规则的。

Google 译文：图 3：相位一致性。为了说明的目的，选择示例在模型之间大致相似。顶行显示波形模数音符的基本周期（MIDI C60），对于在音符中间拍摄的 1028 个示例。请注意，由于波形非常周期性，因此实际数据完全重叠。然而，WaveGAN 和 PhaseGAN 具有许多相位不规则性，从而产生模糊的线条网。 IFGAN 更加连贯，从周期到周期只有很小的变化。在下面的 Rainbowgrams 中，真实数据和 IF 模型具有相干波形，可以为每个谐波产生强一致的颜色，而 PhaseGAN 由于相位不连续而具有许多斑点，并且 WaveGAN 模型非常不规则。

Phase Coherence

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.

百度译文：图 3 显示了来自不同 GaN 变体的示例的相位一致性。从顶部以基频包裹的波形可以清楚地看出，真实数据和 IF 模型产生的波形在各个周期都是一致的。相位根有一些相位不连续，而波干则很不规则。下面我们使用彩虹图（Engel 等人，2017）将频率的对数量级描述为亮度，将 if 描述为彩虹颜色图上的颜色。这种可视化有助于通过强一致的颜色来观察真实数据和 ifgan 中谐波的清晰相位相干。相比之下，相位差的不连续性表现为散斑噪声，而波干在很大程度上是不相干的。

Google 译文：图 3 显示了来自不同 GAN 变体的实例的相位相干性。从顶部的波形可以清楚地看出，它们以基频包裹，真实数据和 IF 模型产生的波形在周期与周期之间是一致的。 PhaseGAN 具有一些相位不连续性，而 WaveGAN 非常不规则。下面我们使用 Rainbowgrams（Engel et al。，2017）来描述频率的对数幅度作为亮度，IF 作为彩虹色图上的颜色。这种可视化有助于通过强一致的颜色看到真实数据和 IFGAN 中谐波的清晰相位一致性。相比之下，PhaseGAN 不连续性表现为斑点噪声，而 WaveGAN 似乎在很大程度上不相干。

INTERPOLATION

百度译文：插值

Google 译文：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. 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

百度译文：正如在引言中所讨论的，gans 还允许在相同的潜在向量上调节整个序列，而不是像 wavenet 这样的记忆密集型自回归模型只允许短的子序列。Wavenet 自动编码器，如（Engel 等人，2017 年）中的，学习控制毫秒级生成但范围有限的本地潜在代码，并且具有自己的结构，必须建模，而不能确定紧凑的先验。在图 4 中，我们采用一个预训练的 wavenet autoencoder 5，比较原始波形（top）、wavenet autoencoder 的分布式潜在代码和 if-mel-gan 的全局代码中的示例之间的插值。插入波形在知觉上等同于两个不同声音的振幅之间的混合。Wavenet 通过在音色空间中混合而改进了这两个音符，但是线性插值不符合复杂的先验延迟，中间音有分解、振荡和鸣笛的趋势，这是 Wavenet 模型的自然失效模式。然而，Gan 模型有一个球形高斯先验，它被解码为

Google 译文：正如在引言中所讨论的，GAN 还允许对整个序列的相同潜在向量进行条件化，而不是像 WaveNet 这样的内存密集型自回归模型的短子序列。 WaveNet 自动编码器，例如（Engel 等，2017）中的那些，学习本地潜在的代码，这些代码控制毫秒级的生成，但范围有限，并且具有自己的结构，必须建模并且不能紧凑之前。在图 4 中，我们采用预训练的 WaveNet 自动编码器 5，并比较原始波形（顶部）中示例之间的插值，WaveNet 自动编码器的分布式潜码以及 IF-Mel GAN 的全局代码。对波形进行插值在感知上等效于两个不同声音的幅度之间的混合。 WaveNet 通过在音色空间中混合来改善这两个音符，但是线性插值与潜伏中的复杂先验不对应，并且中间声音具有分离，振荡和吹哨的倾向，这是自然故障 WaveNet 模型的模式。然而，GAN 模型具有球形高斯先验，其被解码为

nsynth

百度译文：恩辛斯

Google 译文：nsynth

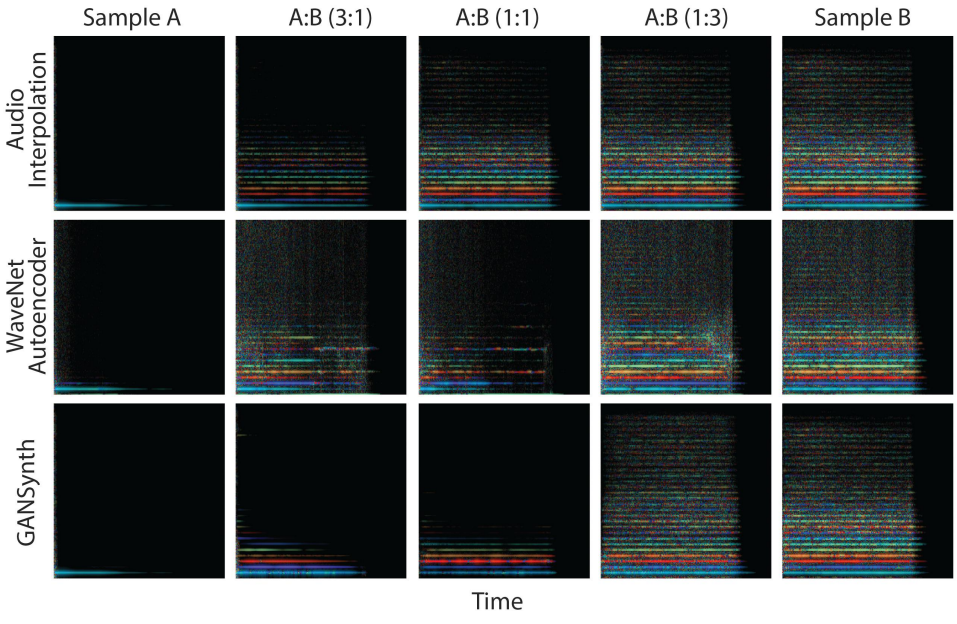

Figure 4: Global interpolation. Examples available for listening4. Interpolating between waveforms perceptually results in crossfading the volumes of two distinct sounds (rainbowgrams at top). The WaveNet autoencoder (middle) only has local conditioning distributed in time, and no compact prior over those time series, so linear interpolation ventures off the true prior / data manifold, and produces in-between sounds that are less realistic examples and feature the default failure mode of autoregressive wavenets (feedback harmonics). Meanwhile, the IF-Mel GAN (bottom) has global conditioning so interpolating in perceptual attributes while staying along the prior at all intermediate points, so they produce high-fidelity audio examples like the endpoints.

百度译文：图 4：全局插值。可供收听的示例 4。在两个波形之间插入感知结果交叉衰落两个不同的声音（彩虹图在顶部）。Wavenet AutoEncoder（Middle）只在时间上具有局部条件分布，在这些时间序列中没有紧凑的先验，因此线性插值冒险离开真正的先验/数据流形，在声音之间产生不太现实的例子，并具有自回归 Wavenet（反馈谐波）的默认故障模式。同时，if-mel-gan（底部）具有全局条件作用，因此可以在所有中间点上沿先验点插入感知属性，从而生成类似端点的高级音频示例。

Google 译文：图 4：全局插值。可用于收听的示例 4。波形之间的插值在感知上导致交叉淡化两个不同声音的音量（顶部的彩虹图）。 WaveNet 自动编码器（中间）仅具有及时分布的局部条件，并且在这些时间序列之前没有紧凑的先验，因此线性插值冒出真正的先前/数据流形，并产生中间声音，这些声音不太真实，并且具有默认值自回归波网的失效模式（反馈谐波）。同时，IF-Mel GAN（底部）具有全局条件，因此在感知属性中进行插值，同时保持在所有中间点的先验，因此它们产生高端音频示例，如端点。

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.

百度译文：产生整个声音，球面插值保持与前面的很好的对齐。因此，插值过程中的感知变化是平滑的，所有中间的潜在向量都被解码以产生声音，而不产生额外的伪影。作为一个更为音乐的例子，在音频例子中，我们在使用 G 大调 6 中巴赫组曲 1 号前奏曲的音调时，在潜在空间中插入 15 个随机点之间的音色。如附录图 7 所示，声音的音色在许多乐器之间平稳地变形，而音高始终跟随所组成的乐章。

Google 译文：产生整个声音，球面插值保持与先前的良好对齐。因此，插值期间的感知变化是平滑的，并且所有中间潜在矢量被解码以产生声音而没有额外的伪像。作为一个更具音乐性的例子，在音频例子中，我们在潜在空间中的 15 个随机点之间插入音色，同时使用 G 大调 6 中 Bach 的 1 号套件前奏的音高。如附录图 7 所示，音色声音在许多乐器之间平滑变形，而音高始终跟随合成乐曲。

Consistent Timbre Across Pitch

While timbre slightly varies for a natural instrument across register, on the whole it remains consis- tent, giving the instrument its unique character. In the audio examples 7, we fix the latent condition- ing 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.

百度译文：虽然自然乐器的音色在整个音域上略有不同，但总体而言，它仍然保持一致性，赋予乐器独特的特性。在音频示例 7 中，我们确定潜在条件变量，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，甘的音色特征基本上保持不变，为潜在空间中的某一点创造了一种独特的乐器特征。如附录图 7 所示，用一个潜在矢量呈现的巴赫前奏曲在一系列音高上具有一致的和声结构。

Google 译文：虽然音色对于寄存器中的自然乐器略有不同，但总体上它仍然是一致的，使乐器具有独特的特性。在音频示例 7 中，我们通过改变五个八度音阶的音调调节来确定潜在的调节变量并生成示例。很明显，GAN 的基本身份基本保持不变，为潜在空间中的给定点创造了独特的仪器身份。如附录图 7 所示，用单个潜在向量渲染的巴赫前奏在一系列音高上具有一致的谐波结构。

Fast Generation

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

百度译文：与自回归模型相比，具有上采样卷积的 GAN 的一个优点是，可以对整个音频样本并行处理训练和生成。这个

Google 译文：GAN 与自回归模型的上采样卷积的优势之一是，可以对整个音频样本并行处理训练和生成。这个

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.

百度译文：非常适合于现代 GPU 硬件，通常 I/O 与迭代自回归算法绑定。当我们在 Titanx GPU 上合成一个 4 秒的音频样本时，可以看到这一点，完成延迟从波网基线的 1077.53 秒下降到 if-mel-gan 的 20 毫秒，使其速度快了大约 53880 倍。以前的 wavenet 自动编码器的应用都是在 nsynth 数据集上训练的，它们的音乐性能依赖于预渲染所有可能的声音以供播放，因为合成延迟很长 8。这项工作为设备上的实时神经网络音频合成开辟了有趣的可能性，允许用户探索更广泛的表达声音的调色板。

Google 译文：非常适合现代 GPU 硬件，它通常与迭代自回归算法进行 I / O 绑定。当我们在 TitanX GPU 上合成单个四秒音频样本时，可以看到这一点，完成延迟从 WaveNet 基线的 1077.53 秒下降到 IF-Mel GAN 的 20 毫秒，使其快大约 53,880 倍。之前应用于 NSynth 数据集的 WaveNet 自动编码器的音乐性能依赖于预渲染所有可能的声音，因为合成延迟时间长 8.这项工作为设备上的实时神经网络音频合成开辟了有趣的可能性，允许用户使用探索更广泛的富有表现力的声音。

Related Work

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. 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. 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. 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 ad- versarial 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 au- toencoders 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.

百度译文：关于音频的深层次生成模型的许多工作往往侧重于语音合成（van den oord 等人，2018 年；Sotelo 等人，2017 年；Wang 等人，2017 年）。这些数据集需要处理可变长度条件（音素/文本）和音频，并且通常依赖于用于可变长度输入和输出的循环和/或自回归模型。将对抗性音频合成与这些方法进行比较是很有意思的，但是我们把这留给了将来的工作，因为调整 gan 以使用可变长度调节或循环发生器是当前工作的重要扩展。与语音相比，音乐的音频生成相对不足。van den Oord 等人（2016）和 Mehri 等人（2017）提出自回归模型，并展示其合成乐器声音的能力，但这些都受到上述缓慢生成的影响。多纳休等人。（2019）首次将 gans 应用于音频生成，结果一致，但未达到自回归似然模型的音频精度。我们的工作也建立在赣文学的多个最新进展的基础上。Gulrajani 等人（2017）提出了对 gans 丧失功能的修正，并证明了训练稳定性和架构鲁棒性的改善。卡拉斯等。（2018a）进一步引入渐进式培训，在课程中学习发电机和鉴别器的连续层次，在有限的培训时间内提高发电质量。他们还提出了一些建筑技巧来进一步提高质量，我们在最佳模型中使用了这些技巧。Nsynth 数据集最初是作为“塞莱巴音频”（Liu et al.，2015；Engel et al.，2017）引入的，并使用 WaveNet 自动编码器在乐器音色之间插入，但采样速度非常慢。摩尔等。（2019）通过整合跨领域的混淆损失来扩展这项工作，以实现各种音频源之间的音色转换。笛福塞兹等人（2018）通过训练基于帧的回归模型，将音调和仪器标签映射到原始波形，在 Wavenet au-toEncoorders 上实现显著的采样加速（~2500x）。他们在对数谱图中考虑单峰似然回归损失，并通过 STFT 进行反向传播，这有助于很好的频率估计，但没有提供学习相位相干性或处理多模分布的激励。他们的体系结构还需要大量的通道，减慢了样本生成和培训。

Google 译文：关于音频深度生成模型的大部分工作倾向于关注语音合成（van den Oord 等，2018; Sotelo 等，2017; Wang 等，2017）。这些数据集需要处理可变长度调节（音素/文本）和音频，并且通常依赖于可变长度输入和输出的循环和/或自回归模型。将对抗性音频合成与这些方法进行比较会很有意思，但我们将此留待将来工作，因为使用可变长度调节或循环发生器来适应 GAN 是当前工作的重要扩展。与语音相比，音乐的音频生成相对未被充分探索。 van den Oord 等。（2016）和 Mehri 等。（2017）提出了自回归模型并展示了它们合成乐器声音的能力，但是这些都受到上述缓慢生成的影响。多纳休等人。（2019）首先将 GAN 应用于具有相干结果的音频生成，但未达到自回归可能性模型的音频质量。我们的工作还建立在 GAN 文献的最新进展上。 Gulrajani 等。（2017）提出了对 GAN 损失函数的修改，并证明了改进的训练稳定性和架构稳健性。卡拉斯等人。（2018a）进一步引入渐进式训练，其中在课程中学习生成器和鉴别器的连续层，从而在训练时间有限的情况下提高生成质量。他们还提出了许多建筑技巧，以进一步提高质量，我们采用最好的模型。 NSynth 数据集首先作为“CelebA of audio”（Liu et al。，2015; Engel et al。，2017）引入，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 等人。（2019）通过加入一个广泛的域混淆损失来扩展这项工

作，以实现各种音频源之间的音色变换。 Defossez 等人。 （2018）通过训练基于帧的回归模型，从音高和乐器标签到原始波形，实现了对 wavenet 自动编码器的显着采样加速（∽2,500x）。他们考虑对数谱图中的单峰似然回归损失，并通过 STFT 反向传播，这得出了良好的频率估计，但没有提供学习相位一致性或处理多模态分布的动机。他们的架构还需要大量的通道，减慢了样本生成和培训的速度。

Conclusion

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.

百度译文：通过仔细控制用于生成建模的音频表示，我们在 nsynth 数据集上使用 gans 演示了高质量音频生成，超过了强 Wavenet 基线的灵活性，同时生成了数万倍的样本。虽然这是 Gans 音频生成的主要进展，但本研究集中于特定的受控数据集，需要进一步的工作来验证和扩展它，使其成为更广泛的信号类别，包括语音和其他类型的自然声音。这项工作还为域传输和其他令人兴奋的应用开辟了可能的途径，以对抗音频损失。对于音频来说，模式崩溃和 GANS 常见的多样性问题也存在，我们将进一步研究如何将敌对损失与编码器结合起来，或者更直接的回归损失，以便更好地捕获完整的数据分布。

Google 译文：通过仔细控制用于生成建模的音频表示，我们在 NSynth 数据集上演示了使用 GAN 的高质量音频生成，超过了强大的 WaveNet 基线的优势，同时生成的样本速度提高了数万倍。虽然这是使用 GAN 进行音频生成的一个重大进步，但本研究的重点是特定的受控数据集，需要进一步的工作来验证并将其扩展到更广泛的信号类别，包括语音和其他类型的自然声音。这项工作还为域名转移和其他激动人心的音频损失应用提供了可能的途径。模式崩溃和 GAN 共同的多样性问题也存在于音频中，我们将其留待进一步研究，以考虑将对抗性损失与编码器相结合或更直接的回归损失，以更好地捕获完整的数据分布。

ACKNOWLEDGMENTS

百度译文：致谢

Google 译文：致谢

We would like to thank Rif A. Saurous and David Berthelot for fruitful discussions and help in reviewing the manuscript.

百度译文：我们要感谢 rif a.saurous 和 david berthelot 进行了富有成效的讨论，并帮助他们审阅了手稿。

Google 译文：我们要感谢 Rif A. Saurous 和 David Berthelot 进行富有成效的讨论并帮助审阅手稿。

REFERENCES Mathieu Andreux and Stephane Mallat. Music generation and transformation with moment matching-scattering inverse networks. In International Society for Music Information Retrevial Conference, 2018. URL http://ismir2018.ircam.fr/doc/pdfs/131_Paper.pdf.

百度译文：参考 Mathieu Andreux 和 Stephane Mallat。基于矩匹配散射逆网络的音乐生成与变换。在国际音乐信息检索协会会议上，2018 年。网址：http://ismir2018.ircam.fr/doc/pdfs/131_paper.pdf。

Google 译文：参考文献 Mathieu Andreux 和 Stephane Mallat。基于矩匹配 - 散射逆网络的音乐生成与变换。在国际音乐信息社会革命会议，2018 年.URL http://ismir2018.ircam.fr/doc/pdfs/131_Paper.pdf。

Martin Arjovsky, Soumith Chintala, and L´eon Bottou. Wasserstein generative adversarial networks. In Doina Precup and Yee Whye Teh (eds.), Proceedings of the 34th International Conference on Machine Learning, volume 70 of Proceedings of Machine Learning Research, pp. 214–223, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR. URL http:// proceedings.mlr.press/v70/arjovsky17a.html.

百度译文：Martin Arjovsky、Soumith Chintala 和 L’eon Bottou。瓦瑟斯坦生成的敌对网络。在 Doina Precup 和 Yee Whye Teh（编辑），《第 34 届机器学习国际会议论文集》，《机器学习研究论文集》第 70 卷，第 214-223 页，澳大利亚悉尼国际会议中心，2017 年 8 月 6-11 日。PMLR 网址：http://proceedings.mlr.press/v70/arjovsky17a.html。

Google 译文：Martin Arjovsky，Soumith Chintala 和 L'Thon Bottou。 Wasserstein 生成对抗网络。在 Doina Precup 和 Yee Whye Teh（编辑），第 34 届国际机器学习会议论文集，机器学习研究会议论文集 70，第 214-223 页，国际会议中心，悉尼，澳大利亚，2017 年 8 月 6 日至 11 日.PMLR。 URL http：// proceedings.mlr.press/v70/arjovsky17a.html。

David Berthelot, Thomas Schumm, and Luke Metz. BEGAN: Boundary equilibrium generative

百度译文：大卫·贝塞洛特、托马斯·舒姆和卢克·梅茨。开始：边界平衡生成

Google 译文：David Berthelot，Thomas Schumm 和 Luke Metz。 BEGAN：边界均衡生成

adversarial networks. arXiv preprint arXiv:1703.10717, 2017.

百度译文：对抗性网络。ARXIV 预印 ARXIV:1703.10717，2017.

Google 译文：对抗性网络。 arXiv preprint arXiv：1703.10717,2017。

Boualem Boashash. Estimating and interpreting the instantaneous frequency of a signal. Proceed-

百度译文：博阿勒姆·博沙什。估计和解释信号的瞬时频率。进行-

Google 译文：Boualem Boashash。估计和解释信号的瞬时频率。继续-

ings of the IEEE, 80(4):540–568, 1992.

百度译文：美国电气与电子工程师协会，80（4）：540–5681992 年。

Google 译文：IEEE，80（4）：540-568,1992。

Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. In International Conference on Learning Representations, 2019. URL https://openreview.net/forum?id=B1xsqj09Fm.

百度译文：安德鲁·布洛克、杰夫·多纳休和凯伦·西蒙尼。大型氮化镓培训，用于高精度自然图像合成。在 2019 年国际学习代表大会上。网址：https://openreview.net/forum？ID= B1XQJ09FM。

Google 译文：Andrew Brock，Jeff Donahue 和 Karen Simonyan。用于高级自然图像合成的大规模 GAN 训练。在国际学习代表会议上，2019。URL https://openreview.net/forum?id=B1xsqj09Fm。

Alexandre Defossez, Neil Zeghidour, Nicolas Usunier, Leon Bottou, and Francis Bach. Sing: Symbol-to-instrument neural generator. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), Advances in Neural Information Processing Systems 31, pp. 9041–9051. Curran Associates, Inc., 2018. URL http://papers.nips.cc/paper/ 8118-sing-symbol-to-instrument-neural-generator.pdf.

百度译文：亚历山大德福塞兹，尼尔·泽吉杜尔，尼古拉斯·乌西尼尔，利昂·博托和弗朗西斯·巴赫。歌唱：象征仪表神经发生器。S.Bengio、H.Wallach、H.Larochelle、K.Grauman、N.Cesa Bianchi 和 R.Garnett（编辑），《神经信息处理系统的进展》，31，第 9041-9051 页。Curran Associates，Inc.，2018 年。网址：http://papers.nips.cc/paper/8118-sing-symbol-to-instrument-neuric-generator.pdf。

Google 译文：Alexandre Defossez，Neil Zeghidour，Nicolas Usunier，Leon Bottou 和 Francis Bach。唱：符号到仪器的神经发生器。在 S.Bengio，H.Wallach，H.Lagochelle，K.Grauman，N.Cesa-Bianchi 和 R.Garnett（eds。），Advances in Neural Information Processing Systems 31，pp.9041-9051。 Curran Associates，Inc.，2018。URL http://papers.nips.cc/paper/ 8118-sing-symbol-to-instrument-neural-generator.pdf。

Sander Dieleman and Benjamin Schrauwen. End-to-end learning for music audio. In ICASSP, pp.

百度译文：Sander Dieleman 和 Benjamin Schrauwen。音乐音频的端到端学习。在 ICASSP，PP。

Google 译文：Sander Dieleman 和 Benjamin Schrauwen。音乐音频的端到端学习。在 ICASSP 中，pp。

Mark Dolson. The phase vocoder: A tutorial. Computer Music Journal, 10(4):14–27, 1986.

百度译文：Mark Dolson。相位声码器：教程。计算机音乐杂志，10（4）：14–27，1986 年。

Google 译文：马克·多尔森阶段声码器：一个教程。计算机音乐杂志，10（4）：14-27,1986。

Chris Donahue, Julian McAuley, and Miller Puckette. Adversarial audio synthesis. In ICLR, 2019.

百度译文：Chris Donahue、Julian McAuley 和 Miller Puckette。对抗性音频合成。在 ICLR，2019。

Google 译文：Chris Donahue，Julian McAuley 和 Miller Puckette。对抗性音频合成。在 ICLR，2019 年。

Jesse Engel, Cinjon Resnick, Adam Roberts, Sander Dieleman, Douglas Eck, Karen Simonyan, and Mohammad Norouzi. Neural audio synthesis of musical notes with WaveNet autoencoders. In ICML, 2017.

百度译文：杰西·恩格尔、辛乔恩·雷斯尼克、亚当·罗伯茨、桑德·迪尔曼、道格拉斯·埃克、卡伦·西蒙尼和穆罕默德·诺鲁齐。用 Wavenet 自动编码器对音符进行神经音频合成。在 ICML，2017。

Google 译文：Jesse Engel，Cinjon Resnick，Adam Roberts，Sander Dieleman，Douglas Eck，Karen Simonyan 和 Mohammad Norouzi。使用 WaveNet 自动编码器进行音符的神经音频合成。在 ICML，2017 年。

Philippe Esling, Axel Chemla-Romeu-Santos, and Adrien Bitton. Bridging audio analysis, perception and synthesis with perceptually-regularized variational timbre spaces. In Proceedings of the 19th International Society for Music Information Retrieval Conference, ISMIR 2018, Paris, France, September 23-27, 2018, pp. 175–181, 2018. URL http://ismir2018.ircam.fr/ doc/pdfs/219_Paper.pdf.

百度译文：Philippe Esling、Axel Chemla Romeu Santos 和 Adrien Bitton。将音频分析、感知和合成与感知规则化的变音色空间连接起来。在第 19 届国际音乐信息检索学会会议记录中，ISMIR 2018，法国巴黎，2018 年 9 月 23 日至 27 日，第 175-181 页，2018 年。网址：http://ismir2018.ircam.fr/doc/pdfs/219_paper.pdf。

Google 译文：Philippe Esling，Axel Chemla-Romeu-Santos 和 Adrien Bitton。通过感知正则化的变分音色空间来桥接音频分析，感知和合成。在第 19 届国际音乐信息检索学会会议论文集，ISMIR 2018，法国巴黎，2018 年 9 月 23 日至 27 日，第 175-181 页，2018 年.URL http://ismir2018.ircam.fr/doc/pdfs/ 219_Paper.pdf。

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, In NIPS, pp. 2672–2680,

百度译文：伊恩·古德费罗，让·波吉特·阿巴迪，梅赫迪·米尔扎，宾·许，大卫·沃德·法利，舍吉尔·奥扎尔，尼普斯，第 2672-2680 页，

Google 译文：Ian Goodfellow，Jean Pouget-Abadie，Mehdi Mirza，Bing Xu，David Warde-Farley，Sherjil Ozair，In NIPS，pp.2672-2680，

Aaron Courville, and Yoshua Bengio. Generative adversarial nets. 2014.

百度译文：Aaron Courville 和 Yoshua Bengio。生成对抗网。2014。

Google 译文：Aaron Courville 和 Yoshua Bengio。生成对抗网。 2014 年

Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Im-

百度译文：Ishaan Gulrajani、Faruk Ahmed、Martin Arjovsky、Vincent Dumoulin 和 Aaron C Courville。IM

Google 译文：Ishaan Gulrajani，Faruk Ahmed，Martin Arjovsky，Vincent Dumoulin 和 Aaron C Courville。的 IM

proved training of Wasserstein GANs. In NIPS, pp. 5767–5777, 2017.

百度译文：证明了对华瑟斯坦甘斯的训练。在 NIPS，第 5767–5777 页，2017 年。

Google 译文：证明了 Wasserstein GAN 的训练。在 NIPS，第 5767-5777 页，2017 年。

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In NIPS, pp. 6626–6637. 2017.

百度译文：Martin Heusel、Hubert Ramsauer、Thomas Unterthiner、Bernhard Nessler 和 Sepp Hochreiter。用二次尺度更新规则训练的 GAN 收敛到局部纳什均衡。在 NIPS 中，第 6626–6637 页。2017。

Google 译文：Martin Heusel，Hubert Ramsauer，Thomas Unterthiner，Bernhard Nessler 和 Sepp Hochreiter。由两个时间尺度更新规则训练的 GAN 收敛于局部纳什均衡。在 NIPS，第 6626-6637 页。 2017 年。

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros.

百度译文：Phillip Isola、Jun Yan Zhu、Tinghui Zhou 和 Alexei A Efros。

Google 译文：Phillip ISO 啦, Jun-y 按 Z 虎, ting 会 Zhou, and Alex EIA EF ROS.

conditional adversarial networks. In CVPR, 2017.

百度译文：有条件的对抗性网络。在 CVPR，2017。

Google 译文：条件对抗网络。在 CVPR，2017 年。

Image-to-image translation with

百度译文：图像到图像的翻译

Google 译文：图像到图像的翻译

Yanghua Jin, Jiakai Zhang, Minjun Li, Yingtao Tian, Huachun Zhu, and Zhihao Fang. Towards the automatic anime characters creation with generative adversarial networks. arXiv preprint arXiv:1708.05509, 2017.

百度译文：金阳华、张嘉凯、李敏君、田颖涛、朱华春、方志浩。走向动画人物自动生成与生成对抗网络。ARXIV 预印 ARXIV:1708.05509，2017.

Google 译文：Yanghua Jin, Jiakai Zhang, Minjun Li, Yingtao Tian, Huachun Zhu, and Zhihao Fang. Towards the automatic anime characters creation with generative adversarial networks. arXiv preprint arXiv:1708.05509, 2017.

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for

百度译文：Tero Karras、Timo Aila、Samuli Laine 和 Jaakko Lehtinen。甘薯的渐进生长

Google 译文：Tero Karras，Timo Aila，Samuli Laine 和 Jaakko Lehtinen。 GAN 的渐进式增长

improved quality, stability, and variation. In ICLR, 2018a.

百度译文：提高了质量、稳定性和变异性。在 ICLR，2018A。

Google 译文：提高质量，稳定性和变化。在 ICLR，2018a。

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative ad- versarial networks. CoRR, abs/1812.04948, 2018b. URL http://arxiv.org/abs/1812. 04948.

百度译文：Tero Karras、Samuli Laine 和 Timo Aila。一种基于样式的生成和转换网络的生成器体系结构。corr，abs/1812.04948，2018b.网址：http://arxiv.org/abs/1812.04948。

Google 译文：Tero Karras，Samuli Laine 和 Timo Aila。基于样式的生成器，用于生成的网络。 CoRR，abs / 1812.04948,2018b。网址 http://arxiv.org/abs/1812。 04948。

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. CoRR,

百度译文：Diederik P.Kingma 和 Jimmy BA。亚当：一种随机优化方法。CoRR

Google 译文：Diederik P. Kingma 和 Jimmy Ba。亚当：随机优化的一种方法。 CORR，

abs/1412.6980, 2014. URL http://arxiv.org/abs/1412.6980.

百度译文：ABS/1412.6980，2014 年。网址：http://arxiv.org/abs/1412.6980。

Google 译文：abs / 1412.6980,2014。URL http://arxiv.org/abs/1412.6980。

Naveen Kodali, Jacob Abernethy, James Hays, and Zsolt Kira. On convergence and stability of gans.

百度译文：Naven Kodali、Jacob Abernethy、James Hays 和 zsolt Kira。论甘斯函数的收敛性和稳定性。

Google 译文：Naveen Kodali，Jacob Abernethy，James Hays 和 Zsolt Kira。论甘斯的收敛性和稳定性。

arXiv preprint arXiv:1705.07215, 2017.

百度译文：ARXIV 预印 ARXIV:1705.07215，2017.

Google 译文：arXiv preprint arXiv：1705.07215,2017。

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild.

百度译文：刘紫薇、罗萍、王晓刚和唐晓欧。深入学习的面孔属性在野外。

Google 译文：刘子伟，罗平，王小刚，唐小鸥。深入学习野外的面部属性。

In Proceedings of International Conference on Computer Vision (ICCV), 2015.

百度译文：2015 年国际计算机视觉会议记录。

Google 译文：“计算机视觉国际会议论文集”（ICCV），2015 年。

Soroush Mehri, Kundan Kumar, Ishaan Gulrajani, Rithesh Kumar, Shubham Jain, Jose Sotelo, Aaron Courville, and Yoshua Bengio. SampleRNN: An unconditional end-to-end neural audio generation model. In ICLR, 2017.

百度译文：Soroush Mehri、Kundan Kumar、Ishaan Gulrajani、Rithesh Kumar、Shubham Jain、Jose Sotelo、Aaron Courville 和 Yoshua Bengio。无条件端到端神经音频生成模型。在 ICLR，2017。

Google 译文：Soroush Mehri，Kundan Kumar，Ishaan Gulrajani，Rithesh Kumar，Shubham Jain，Jose Sotelo，Aaron Courville 和 Yoshua Bengio。 SampleRNN：无条件端到端神经音频生成模型。在 ICLR，2017 年

Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization

百度译文：Takeru Miyato、Toshiki Kataoka、Masanori Koyama 和 Yuichi Yoshida。光谱归一化

Google 译文：take 如 MI 呀 to, to 是看 IK A 套卡, MA 三 O 日 KO 亚麻, Andy UI 吃 yo 十大. spectral normalization

for generative adversarial networks. In ICLR, 2018.

百度译文：用于生成对抗性网络。在 ICLR，2018。

Google 译文：用于生成对抗性网络。在 ICLR，2018 年。

Noam Mor, Lior Wolf, Adam Polyak, and Yaniv Taigman. Autoencoder-based music translation. In International Conference on Learning Representations, 2019. URL https://openreview. net/forum?id=HJGkisCcKm.

百度译文：诺姆·莫、莱奥·沃尔夫、亚当·波利亚克和亚尼夫·泰格曼。基于自动编码器的音乐翻译。在 2019 年国际学习代表大会上。网址：https://openreview。网络/论坛？ID= HJGKISCKKM。

Google 译文：Noam Mor，Lior Wolf，Adam Polyak 和 Yaniv Taigman。基于 Autoencoder 的音乐翻译。在国际学习代表会议上，2019。URL https：// openreview。净/论坛？ID = HJGkisCcKm。

Augustus Odena, Christopher Olah, and Jonathon Shlens. Conditional image synthesis with auxil-

百度译文：奥古斯都·奥德纳、克里斯托弗·奥拉和乔纳森·什伦斯。Auxil 条件图像合成-

Google 译文：Augustus Odena，Christopher Olah 和 Jonathon Shlens。辅助条件图像合成

iary classifier gans. In ICML, 2017.

百度译文：我是一流的公司。在 ICML，2017。

Google 译文：iary classi fi er gans。在 ICML，2017 年。

Tom Le Paine, Pooya Khorrami, Shiyu Chang, Yang Zhang, Prajit Ramachandran, Mark A Hasegawa-Johnson, and Thomas S Huang. Fast WaveNet generation algorithm. arXiv preprint arXiv:1611.09482, 2016.

百度译文：Tom Le Paine、Pooya Khorrami、Shiyu Chang、Yang Zhang、Prajit Ramachandran、Mark A Hasegawa Johnson 和 Thomas S Huang。快速波网生成算法。ARXIV 预印 ARXIV:1611.09482，2016.

Google 译文：Tom Le Paine，Pooya Khorrami，Shiyu Chang，Yang Zhang，Prajit Ramachandran，Mark A Hasegawa-Johnson 和 Thomas S Huang。快速 WaveNet 生成算法。 arXiv preprint arXiv：1611.09482,2016。

Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep

百度译文：亚历克·拉德福德、卢克·梅茨和苏米特·钦塔拉。无监督的代表性学习

Google 译文：Alec Radford，Luke Metz 和 Soumith Chintala。无监督的代表性学习与深度

convolutional generative adversarial networks. In ICLR, 2016.

百度译文：卷积生成的对抗性网络。在 ICLR，2016。

Google 译文：卷积生成对抗网络。在 ICLR，2016 年。

Eitan Richardson and Yair Weiss. On GANs and GMMs. CoRR, abs/1805.12462, 2018. URL

百度译文：伊坦·理查森和耶尔·韦斯。关于 GAN 和 GMM。corr，abs/1805.124622018 年。统一资源定位地址

Google 译文：Eitan Richardson 和 Yair Weiss。关于 GAN 和 GMM。 CoRR，abs / 1805.12462,2018。URL

http://arxiv.org/abs/1805.12462.

百度译文：http://arxiv.org/abs/1805.12462。

Google 译文：HTTP://AR 西 V.org/ABS/1805.12462.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen.

百度译文：Tim Salimans、Ian Goodfellow、Wojciech Zaremba、Vicki Cheung、Alec Radford 和陈曦。

Google 译文：Tim Salimans，Ian Goodfellow，Wojciech Zaremba，Vicki Cheung，Alec Radford 和 Xi Chen。

Improved techniques for training gans. In NIPS, pp. 2234–2242, 2016.

百度译文：改进了培训方法。在 NIPS，第 2234-2242 页，2016 年。

Google 译文：改进训练甘蔗的技术。在 NIPS，第 2234-2242 页，2016 年。

Tim Salimans, Andrej Karpathy, Xi Chen, and Diederik P. Kingma.

百度译文：Tim Salimans、Andrej Karpathy、陈曦和 Diederik P. Kingma。

Google 译文：Tim Salimans，Andrej Karpathy，Xi Chen 和 Diederik P. Kingma。

Improv- ing the pixelcnn with discretized logistic mixture likelihood and other modifications. CoRR, abs/1701.05517, 2017. URL http://arxiv.org/abs/1701.05517.

百度译文：用离散化的逻辑混合可能性和其他方法改进 Pixelcn。corr，abs/1701.05517，2017 年。网址：http://arxiv.org/abs/1701.05517。

Google 译文：使用离散化的逻辑混合可能性和其他修改来改进 pixelcnn。 CoRR，abs / 1701.05517,2017。URL http://arxiv.org/abs/1701.05517。

Pixelcnn++:

百度译文：PixelCNn++：

Google 译文：Pixelcnn ++：

Jose Sotelo, Soroush Mehri, Kundan Kumar, Joao Felipe Santos, Kyle Kastner, Aaron Courville,

百度译文：Jose Sotelo、Soroush Mehri、Kundan Kumar、Joao Felipe Santos、Kyle Kastner、Aaron Courville，

Google 译文：Jose Sotelo，Soroush Mehri，Kundan Kumar，Joao Felipe Santos，Kyle Kastner，Aaron Courville，

and Yoshua Bengio. Char2wav: End-to-end speech synthesis. 2017.

百度译文：还有 Yoshua Bengio。char2wav：端到端语音合成。2017。

Google 译文：和 Yoshua Bengio。 Char2wav：端到端语音合成。 2017 年。

Lucas Theis, A¨aron van den Oord, and Matthias Bethge. A note on the evaluation of generative

百度译文：卢卡斯·泰斯、范登奥尔德和马蒂亚斯·贝奇。关于生成性评价的注记

Google 译文：Lucas Theis，A¨aronvanden Oord 和 Matthias Bethge。关于生成性评价的一个注记

models. In ICLR, 2016.

百度译文：模型。在 ICLR，2016。

Google 译文：楷模。在 ICLR，2016 年。

A¨aron 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. In SSW, pp. 125, 2016.

百度译文：A–Aron van den Oord、Sander Dieleman、Heiga Zen、Karen Simonyan、Oriol Vinyals、Alex Graves、Nal Kalchbrenner、Andrew W Senior 和 Koray Kavukcuoglu。wavenet：原始音频的生成模型。SSW，第 125 页，2016 年。

Google 译文：A¨aronvanden Oord，Sander Dieleman，Heiga Zen，Karen Simonyan，Oriol Vinyals，Alex Graves，Nal Kalchbrenner，Andrew W Senior 和 Koray Kavukcuoglu。 Wavenet：原始音频的生成模型。在 SSW，第 125 页，2016 年。

Aaron van den Oord, Yazhe Li, Igor Babuschkin, Karen Simonyan, Oriol Vinyals, Koray Kavukcuoglu, George van den Driessche, Edward Lockhart, Luis Cobo, Florian Stimberg, Nor- man Casagrande, Dominik Grewe, Seb Noury, Sander Dieleman, Erich Elsen, Nal Kalchbrenner, Heiga Zen, Alex Graves, Helen King, Tom Walters, Dan Belov, and Demis Hassabis. Parallel WaveNet: Fast high-fidelity speech synthesis. In Jennifer Dy and Andreas Krause (eds.), Pro- ceedings of the 35th International Conference on Machine Learning, volume 80 of Proceedings of Machine Learning Research, pp. 3918–3926, Stockholmsmssan, Stockholm Sweden, 10–15 Jul 2018. PMLR. URL http://proceedings.mlr.press/v80/oord18a.html.

百度译文：Aaron van den Oord、Yazhe Li、Igor Babuschkin、Karen Simonyan、Oriol Vinyals、Koray Kavukcuoglu、George van den Driessche、Edward Lockhart、Luis Cobo、Florian Stimberg、Nor-man Casagrande、Dominik Grewe、Seb Noury、Sander Dieleman、Erich Elsen、Nal Kalchbrenner、Heiga Zen、Alex Graves、Helen King、Tom

Walters、Dan Belov 和 Demis Hassabis。平行波网：快速高效的语音合成。Jennifer Dy 和 Andreas Krause（编辑），《第 35 届机器学习国际会议论文集》，《机器学习研究》第 80 卷，第 3918-3926 页，Stockholmsmssan，斯德哥尔摩瑞典，2018 年 7 月 10 日至 15 日。PMLR 网址：http://proceedings.mlr.press/v80/oord18a.html。

Google 译文：Aaron van den Oord，Yazhe Li，Igor Babuschkin，Karen Simonyan，Oriol Vinyals，Koray Kavukcuoglu，George van den Driessche，Edward Lockhart，Luis Cobo，Florian Stimberg，Norman Casagrande，Dominik Grewe，Seb Noury，Sander Dieleman，Erich Elsen ，Nal Kalchbrenner，Heiga Zen，Alex Graves，Helen King，Tom Walters，Dan Belov 和 Demis Hassabis。 Parallel WaveNet：快速高保真语音合成。在 Jennifer Dy 和 Andreas Krause（编辑），第 35 届机器学习国际会议论文集，机器学习研究论文集第 80 卷，第 3918-3926 页，Stockholmsmssan，瑞典斯德哥尔摩，2018 年 7 月 10 日至 15 日.PMLR 。 URL http://proceedings.mlr.press/v80/oord18a.html。

Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, et al. Tacotron: Towards end-to-end speech synthesis. In INTERSPEECH, 2017.

百度译文：王玉轩、RJ 斯凯里·瑞安、戴西·斯坦顿、吴永辉、罗恩·J·韦斯、纳夫德·贾利特、杨宗亨、杨颖、陈志峰、陈珊米·本吉奥等。塔克龙：走向端到端语音合成。在 Interspeech，2017 年。

Google 译文：Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, et al. Tacotron: Towards end-to-end speech synthesis. In INTERSPEECH, 2017.

Lior Wolf, Yaniv Taigman, and Adam Polyak. Unsupervised creation of parameterized avatars.

百度译文：Lior Wolf、Yaniv Taigman 和 Adam Polyak。无监督地创建参数化的虚拟人物。

Google 译文：Lior Wolf，Yaniv Taigman 和 Adam Polyak。无监督创建参数化头像。

CoRR, abs/1704.05693, 2017.

百度译文：corr，abs/1704.05693，2017 年。

Google 译文：科尔，从/ 1704.05693，2017 年。

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation

百度译文：朱军、朴泰松、伊索拉和埃夫罗斯。未配对的图像到图像转换

Google 译文：朱俊妍，Taesung Park，Phillip Isola 和 Alexei A Efros。不成对的图像到图像转换

using cycle-consistent adversarial networks. In ICCV, 2017.

百度译文：使用循环一致的对抗网络。在 ICCV，2017。

Google 译文：使用周期一致的对抗网络。在 ICCV，2017 年。

Zhenyao Zhu, Jesse H Engel, and Awni Y Hannun. Learning multiscale features directly fromwave-

百度译文：朱振耀、恩格尔和汉农。直接从 Wave 学习多尺度特征-

Google 译文：朱振耀，Jesse H Engel 和 Awni Y Hannun。直接从 wave-学习多尺度特征

forms. CoRR, vol. abs/1603.09509, 2016.

百度译文：形式。Corr，Vol.Abs/1603.09509，2016 年。

Google 译文：形式。 CoRR，第一卷 abs / 1603.09509,2016。

A MEASURING DIVERSITY ACROSS GENERATED EXAMPLES

百度译文：通过生成的示例测量多样性

Google 译文：跨越生成实例的测量多样性

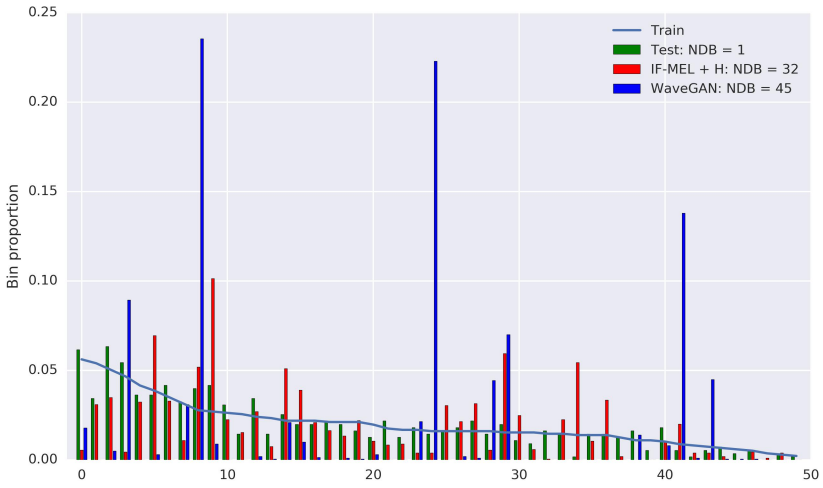

Figure 5: NDB bin proportions for the IF-Mel + H model and the WaveGAN baseline (evaluated with examples of pitch 60).

百度译文：图 5:if-mel+h 模型和 Wavegan 基线的 ndb-bin 比例（以螺距 60 为例进行评估）。

Google 译文：图 5：IF-Mel + H 模型和 WaveGAN 基线的 NDB 柱比例（用间距 60 的实例评估）。

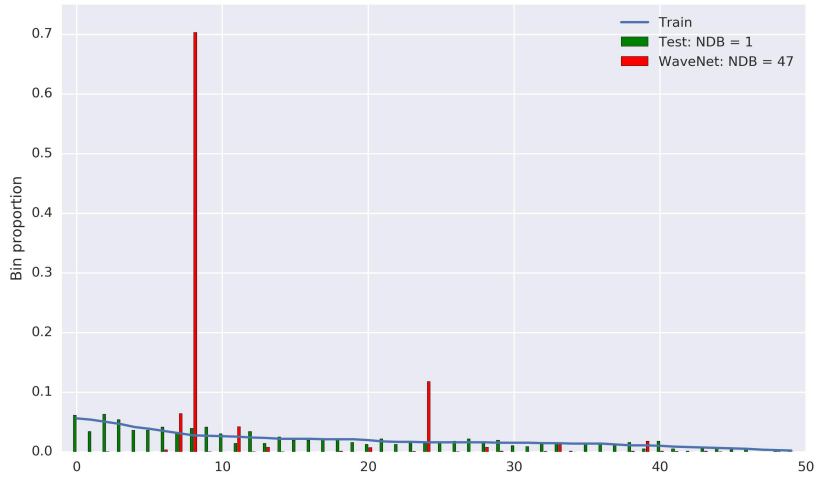

Figure 6: NDB bin proportions for the WaveNet baseline (evaluated with examples of pitch 60).

百度译文：图 6：波网基线的 ndb-bin 比例（以间距 60 为例进行评估）。

Google 译文：图 6：WaveNet 基线的 NDB bin 比例（用间距 60 的示例评估）。

B TIMBRAL SIMILARITY ACROSS PITCH

百度译文：B 音高的音色相似性

Google 译文：B 跨越间距的时间相似性

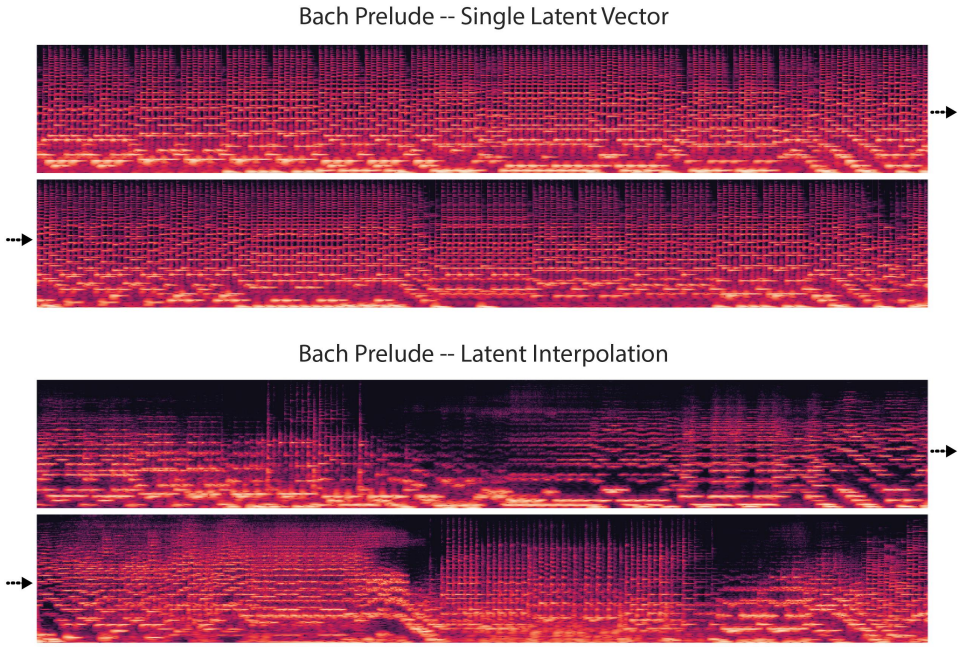

Figure 7: The first 20 seconds (10 seconds per a row) of the prelude to Bach's Suite No. 1 in G major 9, for pitches synthesized with a single latent vector (top), and with spherical interpolation in latent space (bottom). The timbre is constant for a single latent vector, shown by the consistency of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

百度译文：图 7:G 大调 9 巴赫组曲 1 序曲的前 20 秒（每行 10 秒），用一个潜在矢量（上）合成的音高，用潜在空间（下）的球面插值。音色对于单个潜在矢量是恒定的，由上谐波结构的一致性来表示，但随着潜在矢量的变化，音色变化很大。听力示例见 https://goo.gl/magenta/gansynth-examples

Google 译文：图 7：G 大调 9 中巴赫 1 号套房前奏的前 20 秒（每行 10 秒），单个潜在向量合成的音高（上图），潜在空间中的球形插值（下图）。单个潜在向量的音色是恒定的，由高次谐波结构的一致性表示，而随着潜在向量的变化，它会发生显着变化。听力示例在 https://goo.gl/magenta/gansynth-examples 上提供

C BASELINE MODEL COMPARISONS

百度译文：C 基线模型比较

Google 译文：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 outper- forms 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.

百度译文：表 2：直接生成波形的模型比较。我们的波形氮化镓基线形式类似于 Wavegan 基线，但渐进式训练并不能提高性能，因此我们仅将其与本文的 Wavegan 基线进行比较。8 位分类 wavenet 输出器形成了 16 位的物流混合，这可能是由于 16 位模型的稳定性下降，尽管精度有所提高，但仅采用变桨调节。

Google 译文：表 2：直接生成波形的模型的比较。我们的波形 GAN 基线执行类似于 WaveGAN 基线，但渐进式训练不会提高性能，因此我们只与该论文的 WaveGAN 基线进行比较。 8 位分类 WaveNet 优于 16 位物流混合，可能是由于 16 位模型的稳定性降低，只有俯仰调节，尽管增加了灵活性。

Examples WaveGAN Waveform NoProg Waveform Prog WaveNet 8-bit WaveNet 16-bit

百度译文：示例 wavegan 波形 noprog 波形 prog wavenet 8 位 wavenet 16 位

Google 译文：示例 WaveGAN 波形 NoProg 波形编程 WaveNet 8 位 WaveNet 16 位

NDB FID 461 43.0 447 48.2 45.0 375 320 44.8 45.9 656

百度译文：NDB FID 461 43.0 447 48.2 45.0 375 320 44.8 45.9 656

Google 译文：NDB 461 43.0 447 48.2 45.0 375 320 44.8 45.9 656

IS 13.7 14.8 2.5 29.1 9.5

百度译文：IS 13.7 14.8 2.5 29.1 9.5

Google 译文：IS 13.7 14.8 2.5 29.1 9.5

PA 82.7 96.3 56.7 92.7 64.6

百度译文：PA 82.7 96.3 56.7 92.7 64.6

Google 译文：PA 82.7 96.3 56.7 92.7 64.6

PE 1.40 1.61 3.59 0.70 1.71

百度译文：PE 1.40 1.61 3.59 0.70 1.71

Google 译文：PE 1.40 1.61 3.59 0.70 1.71

D TRAINING DETAILS

百度译文：D 培训详情

Google 译文：D 培训细节

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. As in the original progressive GAN paper, both networks use box upscaling/downscaling and the generators use pixel normalization,

百度译文：Gan 架构直接从 Tensor flow 10 中的开源实现中进行了调整。完整的细节在表 3 中给出，包括在 AC-GAN 中为鉴别器的末端添加一个音高等级。所有模型都接受了 Adam 优化器的培训（Kingma&BA，2014 年）。我们对学习率（2e-4、4e-4、8e-4）和辅助分类损失的权重（0.1、1.0、10）进行了扫描，并且对所有变量（光谱表示、渐进/非渐进、频率分辨率）而言，8e-4 和分类损失 10 的学习率表现最佳。与原始的渐进式 GaN 文件一样，两个网络都使用方框上/下缩放，而生成器使用像素标准化，

Google 译文：GAN 架构直接改编自 Tensor 流程 10 中的开源实现。表 3 给出了完整的细节，包括在 AC-GAN 中的鉴别器末尾添加一个音调分类器。所有模型都使用 ADAM 优化器进行训练（Kingma＆Ba，2014）。我们扫描了学习率（2e-4,4e-4,8e-4）和辅助分类器损失的权重（0.1,1.0,10），并找到所有变体（频谱表示，渐进/无渐进，频率分辨率）） 8e-4 的学习率和 10 的分类损失表现最佳。与最初的渐进式 GAN 论文一样，两个网络都使用盒子放大/缩小，并且生成器使用像素标准化，

(cid:88)

百度译文：（CID：88）

Google 译文：（cid：88）

c

百度译文：C

Google 译文：C

x = xnhwc/(

百度译文：x= xNHWC//（

Google 译文：x = xnhwc /（

x2 nhwc)0.5

百度译文：X2-NHWC）0.5

Google 译文：x2 egc）

(1)

百度译文：（1）

Google 译文：(1)

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. 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. 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. 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. 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.

百度译文：其中 n、h、w 和 c 分别表示批次、高度、宽度和通道尺寸，x 表示激活，c 表示通道总数。鉴别器还将小批量激活的标准偏差作为一个标量通道附加到卷积堆栈的末尾，如表 3 所示。由于我们发现对发生器使用 tanh 输出非线性很有帮助，因此在传递给鉴别器之前，我们对实际数据进行了标准化。我们测量超过 100 个例子的最大范围，并独立地移动和缩放对数大小和相位至-0.8，0.8]以允许离群值，并使用 tanh 非线性的更多线性区域。我们在一个 V100 GPU 上对每个 GAN 变体进行为期 4.5 天的培训，批量大小为 8。对于非渐进式模型，这相当于对~5 米示例的培训。对于渐进式模型，我们对每个阶段（7 个阶段）1.6 米的示例进行培训，在阿尔法混合过程中培训 800K，在混合后培训 800K。在最后一个阶段，我们继续培训，直到 4.5 天结束。因为早期的列车速度更快，渐

进式模型在～11m 的例子上列车。对于 wavenet 基线，我们还调整了开源张量流实现 11。解码器由 30 层扩展卷积组成，每个通道 512 个，接收区 3 个，每个卷积跳过 1 个连接到输出端。各层分为 3 层，每层 10 层，膨胀从 20 增加到 29，然后重复。我们用一个调节堆栈替换音频编码器堆栈，该堆栈在一个按时间分布的热节距调节信号上运行（3 秒打开，1 秒关闭）。调节栈是 5 层扩张卷积，增加到 25 层，然后是 3 层规则卷积，都有 512 个通道。然后，该调节信号通过解码器每层的 1x1 卷积，并添加到每层的输出，就像在 WaveNet 调节的其他实现中一样。对于 8 位模型，我们使用音频的 mu-律编码和分类丢失，而对于 16 位模型，我们使用 10 个物流的量化混合（Salimans 等人，2017）。Wavenets 在 2 天内收敛到 150k 次迭代，32 个 V100 GPU 接受同步 SGD 培训，每个 GPU 的批量大小为 1，总批量大小为 32。

Google 译文：其中 n，h，w 和 c 分别表示批次，高度，宽度和通道尺寸，x 是激活，C 是通道总数。鉴别器还将小批量激活的标准偏差附加为卷积堆栈末端附近的标量通道，如表 3 所示。由于我们发现使用 Tanh 输出非线性有助于生成器，我们在传递给实际数据之前将其标准化。鉴别者。我们测量超过 100 个示例的最大范围，并且无差别地将对数幅度和相位移位并缩放到[-0.8,0.8]以允许异常值并使用更多 Tanh 非线性的线性方案。我们在单个 V100 GPU 上训练每个 GAN 变体 4.5 天，批量大小为 8.对于非渐进式模型，这相当于~5M 示例的训练。对于渐进模型，我们每阶段（7 个阶段）训练 1.6M 示例，在 alpha 混合期间训练 800k，在混合后训练 800k。在最后阶段，我们继续训练直到 4.5 天完成。因为早期阶段训练更快，渐进模型训练~11M 例子。对于 WaveNet 基线，我们还调整了开源 Tensor 流程实现 11.解码器由 30 层扩张卷积组成，每个层包含 512 个通道和 3 个接收域，每个都有一个 1x1 卷积跳过连接到输出。将这些层分成 3 个 10 个堆叠，每个堆叠中的扩张从 20 增加到 29，然后重复。我们用一个调节堆替换音频编码器堆栈，该调节堆操作在一个时间上分配的单热调节调节信号上（3 秒开，1 秒关）。调节叠层是 5 层扩张卷积，增加到 25 层，然后是 3 层常规卷积，全部有 512 个通道。然后，该调节信号通过解码器的每层的 1x1 卷积并且被添加到每层的输出，如在 WaveNet 调节的其他实施方式中那样。对于 8 位模型，我们使用音频的 mu-law 编码和分类丢失，而对于 16 位模型，我们使用 10 物流的量化混合（Salimans 等，2017）。 WaveNets 在 2 天内融合至 150k 次迭代，32 个 V100 GPU 采用同步 SGD 培训，每个 GPU 批量大小为 1，总批量为 32。

progressive_gan

百度译文：进步主义

Google 译文：progressive_gan

nsynth

百度译文：恩辛斯

Google 译文：nsynth

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.

百度译文：表 3：高频分辨率的模型体系结构。低频分辨率从宽度 4 开始，高度 8，但在其他方面是相同的。“pn”代表像素范数，“lrelu”代表泄漏的直线单位，斜率为 0.2。潜在矢量 Z 有 256 个维度，音高调节是 61 个维度的一个热矢量。

Google 译文：表 3：高频分辨率的模型架构。低频分辨率的宽度为 4，高度为 8，但其他方面相同。 “PN”代表像素范数，“LReLU”代表泄漏的整流线性单位，斜率为 0.2。潜在矢量 Z 具有 256 维，并且间距调节是 61 维单热矢量。

kF ilters Nonlinearity

百度译文：KF 滤波器非线性

Google 译文：kF ilters 非线性

Generator concat(Z, Pitch) conv2d conv2d upsample 2x2 conv2d conv2d upsample 2x2 conv2d conv2d upsample 2x2 conv2d conv2d upsample 2x2 conv2d conv2d upsample 2x2 conv2d conv2d upsample 2x2 conv2d conv2d generator output Discriminator image conv2d conv2d conv2d downsample 2x2 conv2d conv2d downsample 2x2 conv2d conv2d downsample 2x2 conv2d conv2d downsample 2x2 conv2d conv2d downsample 2x2 conv2d conv2d downsample 2x2 concat(x, minibatch std.) conv2d conv2d pitch classifier discriminator output

百度译文：生成器 concat（z，pitch）conv2d conv2d 上采样 2x2 conv2d 上采样 2x2 conv2d 上采样 2x2 conv2d 上采样 2x2 conv2d 上采样 2x2 conv2d 上采样 2x2 conv2d 上采样 2x2 conv2d 生成器输出鉴别器图像 conv2d conv2d 下采样 2x2 conv2d 下采样 2x2 conv2d 下采样 2x2 conv2d 下采样 2x2 conv2d 下采样 2x2 conv2d 示例 2X2 conv2d conv2d downsample 2X2 conv2d downsample 2X2 concat（x，minibatch std.）conv2d conv2d pitch classi fier 鉴别器输出

Google 译文：发生器 concat（Z，Pitch）conv2d conv2d 上采样 2x2 conv2d conv2d 上采样 2x2 conv2d conv2d 上采样 2x2 conv2d conv2d 上采样 2x2 conv2d conv2d 上采样 2x2 conv2d conv2d 上采样 2x2 conv2d conv2d 发生器输出判别器图像 conv2d conv2d conv2d 下采样 2x2 conv2d conv2d 下采样 2x2 conv2d conv2d 下采样 2x2 conv2d conv2d 下采样 2x2 conv2d conv2d 下采样 2x2 conv2d conv2d 下采样 2x2 concat（x，minibatch std。） conv2d conv2d pitch classi fi er discriminator output

Output Size (1, 1, 317) (2, 16, 256) (2, 16, 256) (4, 32, 256) (4, 32, 256) (4, 32, 256) (8, 64, 256) (8, 64, 256) (8, 64, 256) (16, 128, 256) (16, 128, 256) (16, 128, 256) (32, 256, 256) (32,

256, 128) (32, 256, 128) (64, 512, 128) (64, 512, 64) (64, 512, 64) (128, 1024, 64) (128, 1024, 32) (128, 1024, 32) (128, 1024, 2)

百度译文：输出大小（1，1，317）（2，16，256）（2，16，256）（4，32，256）（4，32，256）（4，32，256）（8，64，256）（8，64，256）（8，64，256）（16，128，256）（16，128，256）（16，128，256）（32，256）（32，256，128）（32，256，128）（64，512，128）（64，512，64）（64，512，64）（128，1024，64）（128，1024，32）（128，1024，32）（128，1024，32）（128，1024，2）

Google 译文：输出大小（1,1,327）（2,16,256）（2,16,256）（4,32,256）（4,32,256）（4,32,256）（8,64,256））（8,64,256）（8,124,256）（16,128,256）（16,128,256）（16,128,256）（32,256,256）（32,256,128）（32,512,128）（64,512,128）（64,512,64）（64,512,64）（128,1024,64）（128,1024,32）（128,1024,32）（128,1024,2）

(128, 1024, 2) (128, 1024, 32) (128, 1024, 32) (128, 1024, 32) (64, 512, 32) (64, 512, 64) (64, 512, 64) (32, 256, 64) (32, 256, 128) (32, 256, 128) (16, 128, 128) (16, 128, 256) (16, 128, 256) (8, 64, 256) (8, 64, 256) (8, 64, 256) (4, 32, 256) (4, 32, 256) (4, 32, 256) (2, 16, 256) (2, 16, 257) (2, 16, 256) (2, 16, 256) (1, 1, 61) (1, 1, 1)

百度译文：（128、1024、2）（128、1024、2）（128、1024、2）（128、1024、32）（64、512、32）（64、512、32）（64、512、64）（64、512、64）（64、512、64）（32、256、128）（32、256、128）（32、256、128）（16、128、128、128、128）（16、128、128、256）（（16、128、128、128、128、24、2）（128、1024、128、1024、32、10244、32、256）（128、1024、24、32、32）（128）（128（128、1024、1024、32）（32）（128）（128、1024）（128、1024）（128、1024、1024、1024（1，1，1）

Google 译文：(128, 1024, 2) (128, 1024, 32) (128, 1024, 32) (128, 1024, 32) (64, 512, 32) (64, 512, 64) (64, 512, 64) (32, 256, 64) (32, 256, 128) (32, 256, 128) (16, 128, 128) (16, 128, 256) (16, 128, 256) (8, 64, 256) (8, 64, 256) (8, 64, 256) (4, 32, 256) (4, 32, 256) (4, 32, 256) (2, 16, 256) (2, 16, 257) (2, 16, 256) (2, 16, 256) (1, 1, 61) (1, 1, 1)

kW idth

百度译文：千瓦 IDH

Google 译文：kW 宽度

- 2 3 - 3 3 - 3 3 - 3 3 - 3 3 - 3 3 - 3 3 1

百度译文：-2 3-3 3-3 3-3 3-3 3-3 3 3-3 3 3 1

Google 译文：- 2 3 - 3 3 - 3 3 - 3 3 - 3 3 - 3 3 - 3 3 1

kHeight

百度译文：克高

Google 译文：kHeight

- 16 3 - 3 3 - 3 3 - 3 3 - 3 3 - 3 3 - 3 3 1

百度译文：-16 3-3 3-3 3-3 3-3 3-3 3 3-3 3 3 1

Google 译文：- 16 3 - 3 3 - 3 3 - 3 3 - 3 3 - 3 3 - 3 3 1

- 256 256 - 256 256 - 256 256 - 256 256 - 128 128 - 64 64 - 32 32 2

百度译文：-256 256-256 256-256 256-256 256-128 128-64 64-32 32 2

Google 译文：- 256 256 - 256 256 - 256 256 - 256 256 - 128 128 - 64 64 - 32 32 2

- 1 3 3 - 3 3 - 3 3 - 3 3 - 3 3 - 3 3 - - 3 3 - -

百度译文：-1 3 3-3 3-3 3-3 3-3 3-3 3-3 3-3 3 3--

Google 译文：- 1 3 3 - 3 3 - 3 3 - 3 3 - 3 3 - 3 3 - - 3 3 - -

- 32 32 32 - 64 64 - 128 128 - 256 256 - 256 256 - 256 256 - - 256 256 61 1

百度译文：-32 32 32 32-64 64-128 128-256 256-256 256-256 256-256 256-256 256-256 256 61 1

Google 译文：- 32 32 32 - 64 64 - 128 128 - 256 256 - 256 256 - 256 256 - - 256 256 61 1

- 1 3 3 - 3 3 - 3 3 - 3 3 - 3 3 - 3 3 - - 3 3 - -

百度译文：-1 3 3-3 3-3 3-3 3-3 3-3 3-3 3-3 3 3--

Google 译文：- 1 3 3 - 3 3 - 3 3 - 3 3 - 3 3 - 3 3 - - 3 3 - -

-

百度译文：-

Google 译文：-

PN(LReLU) PN(LReLU)

百度译文：pn（lrelu）pn（lrelu）

Google 译文：PN（LReLU）\ t

-

百度译文：-

Google 译文：-

PN(LReLU) PN(LReLU)

百度译文：pn（lrelu）pn（lrelu）

Google 译文：PN（LReLU）\ t

-

百度译文：-

Google 译文：-

PN(LReLU) PN(LReLU)

百度译文：pn（lrelu）pn（lrelu）

Google 译文：PN（LReLU）\ t

-

百度译文：-

Google 译文：-

PN(LReLU) PN(LReLU)

百度译文：pn（lrelu）pn（lrelu）

Google 译文：PN（LReLU）\ t

-

百度译文：-

Google 译文：-

PN(LReLU) PN(LReLU)

百度译文：pn（lrelu）pn（lrelu）

Google 译文：PN（LReLU）\ t

-

百度译文：-

Google 译文：-

PN(LReLU) PN(LReLU)

百度译文：pn（lrelu）pn（lrelu）

Google 译文：PN（LReLU）\ t

-

百度译文：-

Google 译文：-

PN(LReLU) PN(LReLU)

百度译文：pn（lrelu）pn（lrelu）

Google 译文：PN（LReLU）\ t

Tanh

百度译文：坦赫

Google 译文：正切

- -

百度译文：-

Google 译文：- -

LReLU LReLU

百度译文：勒勒鲁

Google 译文：LRELU LReLU

-

百度译文：-

Google 译文：-

LReLU LReLU

百度译文：勒勒鲁

Google 译文：LRELU LReLU

-

百度译文：-

Google 译文：-

LReLU LReLU

百度译文：勒勒鲁

Google 译文：LRELU LReLU

-

百度译文：-

Google 译文：-

LReLU LReLU

百度译文：勒勒鲁

Google 译文：LRELU LReLU

-

百度译文：-

Google 译文：-

LReLU LReLU

百度译文：勒勒鲁

Google 译文：LRELU LReLU

-

百度译文：-

Google 译文：-

LReLU LReLU

百度译文：勒勒鲁

Google 译文：LRELU LReLU

- -

百度译文：-

Google 译文：- -

LReLU LReLU Softmax

百度译文：LRELU LRELU 软最大值

Google 译文：LReLU LReLU Softmax

-

百度译文：-

Google 译文：-