DARKGAN: EXPLOITING KNOWLEDGE DISTILLATION FOR COMPREHENSIBLE AUDIO SYNTHESIS WITH GANS

Javier Nistal Hurlé, Stefan Lattner, Gael Richard

DARKGAN: EXPLOITING KNOWLEDGE DISTILLATION FOR COMPREHENSIBLE AUDIO SYNTHESIS WITH GANS

Javier Nistal,[1,2] **Stefan Lattner,**[2] **Gaël Richard**[1]
[1]LTCI, Telecom Paris, IP Paris, France
[2]Sony Computer Science Laboratories (CSL), Paris, France

ABSTRACT

Generative Adversarial Networks (GANs) have achieved excellent audio synthesis quality in the last years. However, making them operable with semantically meaningful controls remains an open challenge. An obvious approach is to control the GAN by conditioning it on metadata contained in audio datasets. Unfortunately, audio datasets often lack the desired annotations, especially in the musical domain. A way to circumvent this lack of annotations is to generate them, for example, with an automatic audio-tagging system. The output probabilities of such systems (so-called "soft labels") carry rich information about the characteristics of the respective audios and can be used to distill the knowledge from a teacher model into a student model. In this work, we perform knowledge distillation from a large audio tagging system into an adversarial audio synthesizer that we call DarkGAN. Results show that DarkGAN can synthesize musical audio with acceptable quality and exhibits moderate attribute control even with out-of-distribution input conditioning. We release the code and provide audio examples on the accompanying website.

1. INTRODUCTION

Generative Adversarial Networks (GANs) [1] have achieved impressive results in image and audio synthesis [2–6]. However, it is still an open challenge to learn comprehensible features that capture semantically meaningful properties of the data. In the graphical domain, semantic control is achieved with GANs using semantic layouts [4] or high-level attributes learned through unsupervised methods [2]. Other works achieve disentanglement through regularization terms [7] or explore the latent space for human-interpretable factors of variation [8, 9]. The great success of these approaches is partly enabled by the availability of large-scale image datasets containing rich semantic annotations [10–12]. However, the context is different in the audio domain, where datasets are scarce and often limited in size and availability of annotations.

Therefore, in this work, we test if limited annotations in audio datasets can be circumvented by taking a Knowledge Distillation (KD) approach. To that end, we utilize the soft labels generated by a pre-trained audio-tagging system for conditioning a GAN in an audio generation task. More precisely, we train the GAN on a subset of the NSynth dataset [13], which contains a wide range of instruments from acoustic, electronic, and synthetic sources. For that dataset we generate soft labels with a publicly available audio-tagging model [14], pre-trained with attributes of the AudioSet ontology [15]. This ontology contains a structured collection of sound events from many different sources and descriptions of around 600 attributes obtained from YouTube videos (e.g., "singing bowl", "sonar", "car", "siren", or "bird").

The soft labels indicate how much of the different characteristics are contained in a specific sound (e.g., a synthesizer sound may have some similarity with a singing bowl or a sonar pulse). We hope that the generative model can distill such characteristics (e.g., the "essence" of a singing bowl sound) and is then able to emphasize them in the generation. The slight similarities to specific categories in data that can be distilled using soft labels were coined "Dark Knowledge" in [16]. Therefore, we call the proposed model DarkGAN.

This paper introduces a generic audio cross-task KD framework for transferring semantically meaningful features into a neural audio synthesizer. We implement this framework in DarkGAN, an adversarial audio synthesizer for comprehensible and controllable audio synthesis. We perform an experimental evaluation on the quality of the generated material and the semantic consistency of the learned attribute controls. Numerous audio examples are provided in the accompanying web page, [1] and the code is released for reproducibility. [2]

In what follows, we first mention relevant state-of-the-art works in neural audio synthesis and KD (see Sec. 2). In Sec. 3, some background on dark knowledge and KD is given, and its application to controllable neural audio synthesis is motivated. Next, we describe the experimental framework of DarkGAN (see Sec. 4). In Sec. 5 we provide a discussion of the results, and conclude in Sec. 6.

[1] https://an-1673.github.io/DarkGAN.io/
[2] https://github.com/SonyCSLParis/DarkGAN

2. PREVIOUS WORK

In this section we review some of the most important works on neural audio synthesis and knowledge distillation, paying particular attention to those works tackling tasks similar to ours.

2.1 Neural Audio Synthesis

Many works have applied deep generative methods to address general audio synthesis. These can be categorised into *exact*, *approximate*, and *implicit* density estimation methods. In the first category, autoregressive models of raw audio are state-of-the-art in different audio synthesis tasks [13, 17, 18]. Popular *approximate* density estimation methods are based on Variational Auto-Encoders (VAE) [19]. One of the main advantages of VAEs compared to other approaches is the control they offer over the generative process by manipulating a latent space learned directly from the audio data [20]. Even though latent spaces tend to self-organize according to high-level dependencies in the data, these are still difficult to interpret. Therefore, some works try to impose musically meaningful priors over the structure of these spaces [21–23], or enforce an information bottle-neck by restricting such latent codes to discrete representations to capture fundamental and meaningful features [24, 25].

Generative Adversarial Networks (GANs) [1] belong to the *implicit* density estimation methods. Applications of GANs to audio synthesis have mainly focused on speech tasks [26–32]. The first application to synthesis of musical audio was WaveGAN [33]. Although it did not match autoregressive baselines such as WaveNet [17] in terms of audio quality, it could generate piano and drum sounds quickly and in an entirely unconditional way. Recent improvements in the stabilization and training of GANs [34–36] enabled GANSynth [5] to outperform WaveNet baselines on the task of audio synthesis of musical notes using sparse, pitch conditioning labels. Follow-up works building on GANSynth applied similar architectures to conditional drum sound synthesis using different metadata [6, 37]. DrumGAN [6] synthesizes a variety of drum sounds based on high-level input features describing timbre (e.g., boominess, roughness, sharpness). A few other works have used GANs in a variety of audio tasks like Mel-spectrogram inversion [31], audio domain adaptation [38, 39] or audio enhancement [40].

2.2 Knowledge Distillation

High-performing models are often built upon classifier ensembles that aggregate their predictions to improve the overall accuracy. Despite having excellent performance, these models tend to be large and slow, impeding their use in memory-limited and real-time environments. Different methods exist for optimizing memory consumption and reducing the size of large models or ensembles, e.g., pruning, transfer learning, or quantization. Model compression allows to transfer the function learned by a teacher ensemble or a single large discriminative model into a compact, faster student model exhibiting comparable performance [41]. Instead of training the student model directly on a hand-labeled categorical dataset, this method employs a pre-trained teacher model to re-label the dataset and then train the compact neural network on this teacher-labeled dataset, using the raw predictions as the target. This training framework was shown to yield efficient models which perform better than if they had been trained on the hand-labeled dataset in a variety of discriminative tasks [41–43]. Model compression was further extended and formalized into the general Knowledge Distillation (KD) framework [16].

KD has been extensively applied in various fields and with other ends than model compression [44–46]. An interesting line of research that is closely related to ours proposes cross-task KD from image captioning and classification systems into an image synthesis generative neural-network [46, 47]. In audio, KD was extensively used on Automatic Speech Recognition (ASR) tasks in order to exploit large unlabelled datasets [43], distill the knowledge from deep Recurrent Neural Networks (RNN) [48] or, inversely, to improve the performance of deep RNN models by distilling knowledge from simple models as a regularization technique [49]. Works related to ours use KD as a means to adapt a model to a different audio domain task [50] or even data modality (by distilling knowledge from a video classifier) [51], where labeled datasets are scarce, and large models would easily overfit. Some works employ KD to fuse knowledge from different audio representations into a single compact model [52].

3. BACKGROUND

This section provides a brief introduction to dark knowledge and explains the general knowledge distillation framework.

3.1 Dark Knowledge

In the seminal work on Knowledge Distillation (KD) [16], the authors demonstrate that the improved performance of smaller models is due to the implicit information existent in the teacher's output probabilities (i.e., soft labels). As opposed to hard labels, soft labels contain probability values for all of the output classes. The relative probability values that a specific data instance takes for each class contain information about how the teacher generalized the discriminative task. This hidden information existent in the relative probability values was termed *dark knowledge* [53]. An interesting observation on dark knowledge is that in KD, the student model can learn to correctly classify categories even if the training set does not contain examples thereof [16]. In DarkGAN, we test if this principle can be transferred to audio generation. Many of the AudioSet attributes are not directly linked with the actual training data (e.g., the attributes "reverberation", "meow", or "drum" have little or no relationship to the tonal instrument sounds of the NSynth dataset). However, we hope that the implicit dark knowledge existent in the teacher-labeled data can help DarkGAN learn a coherent feature control over such attributes.

3.2 Knowledge Distillation

Multi-label classifiers typically produce a probability distribution over a set of classes by using a *sigmoid* output layer that converts the so-called logit (the NN output before the activation function), z_i, computed for the ith class into a probability q_i as

$$q_i = \frac{1}{1 + e^{-\frac{z_i}{T}}}, \quad (1)$$

where T is a temperature that is typically set to 1. In KD, knowledge is transferred to the distilled model by training it on the teacher-labeled data, using a higher temperature. By that, the distribution gets "compressed," emphasizing lower probability values. The same (higher) temperature is used while training the distilled model, but the temperature is set back to 1 after training. As for cost function, the binary cross-entropy is used as

$$H_s(q) = -\frac{1}{N}\sum_{i=1}^{N} p_i \log(q_i) + (1 - p_i)\log(1 - q_i), \quad (2)$$

where $N = 128$ is the number of attributes, p_i are the soft-labels predicted by the teacher, and q_i is the probability predicted by the student model for the ith class.

4. EXPERIMENTS

In this section, details are given about the conducted experiments. We describe the AudioSet ontology, the teacher and student architectures, the metrics employed for evaluation, and the baselines used for comparison.

4.1 Dataset

We employ a subset from NSynth [13] for our experiments. NSynth contains approximately 300k single-note audios played by more than 1k different instruments from 10 different families. Each sample's onset occurs at time 0. The dataset contains various labels (e.g., pitch, velocity, instrument type), but we only use (i.e., condition the model on) pitch information in this work. Each sample is four seconds long, with a 16kHz sample rate. For computational simplicity, we use only the first second of each sample. Also, we only consider samples with a MIDI pitch range from 44 to 70 (103.83 - 466.16 Hz), resulting in a subset of approximately 90k sounds equally distributed across the pitch classes. For the evaluation, we perform a 90/10% split of the data.

Previous works on adversarial audio synthesis [5, 54] demonstrated that the Magnitude and Instantaneous Frequency of the STFT works well as a representation for harmonic sounds. We use an FFT size of 2048 bins and an overlap of 75%.

4.2 The AudioSet Ontology

AudioSet [15] is a large-scale dataset containing audio data and an ontology of sound events that seek to describe real-world sounds. It was created to set a benchmark in the development of automatic audio event recognition systems, similar to those in computer-vision, such as ImageNet [10]. The dataset consists of a structured vocabulary of 632 audio event classes and a collection of approximately 2M human-labeled 10-second sound clips drawn from YouTube videos. The ontology is specified as a hierarchy of categories with a maximum depth of 6 levels, covering a wide range of human and animal sounds, musical genres and instruments, and environmental sounds. We encourage the reader to visit the corresponding website for a complete description of the ontology. [3]

In this work, we do not employ all of the AudioSet attributes, as many of them refer to properties that are too vague for musical sounds or describe broader time-scale aspects of the sound (e.g., music, chatter, sound effect). Instead, we rank the attributes based on the geometric mean of their 90^{th} percentile (calculated on the predicted class probabilities for each attribute across the dataset), and the teacher's reported accuracy as $\sqrt{p^i_{90_{th}} \times acc^i}$. Then, we take the first 128 attributes according to this ranking.

4.3 Models

In the following, we introduce the teacher model and DarkGAN's architecture.

4.3.1 Pre-trained AudioSet Classifier

In this work, we distill the knowledge from a pre-trained audio-tagging neural network (PANN) trained on raw audio recordings from the AudioSet collection [14]. PANNs were originally proposed for transferring knowledge to other audio pattern recognition tasks. However, we use them to transfer the knowledge to a generative model and steer the generation process through a comprehensible vocabulary of attributes.

We employ the *CNN-14* model from the PANNs [14]. *CNN-14* is built upon a stack of 6 convolution-based blocks containing 2 CNN layers with a kernel size of 3x3. Batch Normalization is applied after every convolutional layer, and a ReLU non-linearity is used as the activation function. After each convolutional block, they apply an average-pooling layer of size 2x2 for down-sampling. Global pooling is applied after the last convolutional layer to summarize the feature maps into a fixed-length vector. An extra fully-connected layer is added to extract embedding features before the output Sigmoid activation function. For more details on the architecture, please refer to [14].

4.3.2 DarkGAN

The proposed GAN architecture, illustrated in Fig. 1, follows the architecture of DrumGAN [6]. The input to G is a concatenation of 128 teacher-labeled AudioSet attributes $\alpha \in [0, 1]^{128}$ (see Sec. 4.2), a one-hot vector $p \in \{0, 1\}^{26}$ containing the pitch class, and a random vector $z \sim \mathcal{N}_{32}(0, 1)$. The resulting vector is placed as a column in the middle of a 4D tensor with $128 + 32 + 26$ convolutional maps. Then, it is fed through a stack of convolutional and box up-sampling blocks to generate the out-

[3] research.google.com/audioset/ontology/

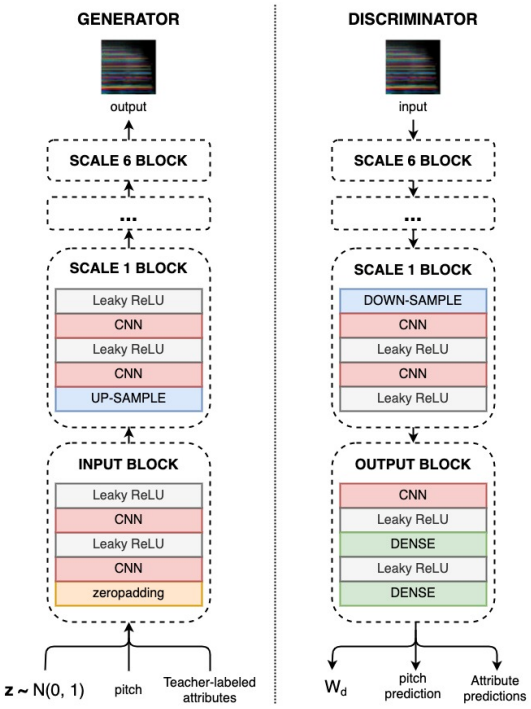

Figure 1. Proposed architecture for DarkGAN [6]

put signal $x = G_\theta(z, p, \alpha)$. The number of feature maps decreases from low to high resolution as {256, 128, 128, 128, 128, 64}. The discriminator D mirrors G's configuration and estimates the Wasserstein distance W_d between the real and generated distributions [35], and predicts the AudioSet features accompanying the input audio in the case of a real batch, or those used for conditioning in the case of generated audio. In order to promote the usage of the conditioning information by G, we add to the objective function an auxiliary binary cross-entropy loss term for the distillation task and a categorical cross-entropy for the pitch classification task [55].

4.4 Evaluation

The task of synthesizing perceptually realistic audio is hard to formalize. In conditional models, as is the case in this work, an additional challenge is to assess whether the model is soundly responsive to the conditional input. In order to evaluate these properties of our model, a diverse set of objective metrics are computed. We compute these metrics for DarkGAN when trained under different temperature values in the distillation process (see Sec. 2.2), as well as for various baselines. In this section, we describe these metrics as well as the baselines used for comparison.

4.4.1 Scores and distances

Following previous methodology [6, 54, 56], we compare real and generated distributions employing these metrics:

- **Inception Score (IS)** [36] penalizes models whose samples cannot be reliably classified into a single class or that only belong to a few from all possible classes. We report on the Pitch Inception Score (PIS) and the Instrument Inception Score (IIS) [54].
- **Kernel Inception Distance (KID)** [57] measures the dissimilarity between embeddings of real and generated samples. A low KID means that the generated and real distributions are close to each other.
- **Fréchet Audio Distance (FAD)** [58] measures the distance between continuous multivariate Gaussians fitted to embeddings of real and generated data. The lower the FAD, the smaller the distance between distributions of real and generated data.

4.4.2 Consistency of Attribute Controls

This work aims to learn semantically meaningful controls with DarkGAN by distilling knowledge from an audio-tagging system trained on attributes from the AudioSet ontology. Therefore, we evaluate if changing an input attribute is reflected in the corresponding output of DarkGAN. To that end, we examine the change in the prediction of the teacher model (w.r.t. the output of DarkGAN) when changing a particular DarkGAN input attribute. A second property to assess is whether the *dark knowledge* helps DarkGAN learn well-formed representations of specific attributes and generalize to out-of-distribution input combinations. To assess these two aspects, we perform the following tests:

1. *Attribute correlation*: we generate 10k samples using attribute vectors from the validation set as input to DarkGAN. The generated samples are fed to the teacher model to predict the attributes again. Then, for each attribute i, we compute the correlation between the input vector α and the predictions $\hat{\alpha}$ as

$$\rho^i(\hat{\alpha}, \alpha) = \rho(F^i(G(z, p, \alpha)), \alpha),$$

where F^i is the classifier's prediction for the ith attribute, p is the pitch, and z is the random noise.

2. *Out-of-distribution Attribute Correlation*: for each attribute i exhibiting a positive correlation, i.e., $\mathcal{S} = \{\rho^i : \rho^i > 0\}$, test (1) is repeated 50 times, but using 1k samples instead of 10k. In each repetition, a specific attribute is progressively incremented by an amount $\delta_l := 10^{-3+l\frac{3.6}{50}}, l = 0, 1, ..., 50$* and we calculate

$$\overline{\rho}_{\delta_l} = \frac{1}{\mid \mathcal{S} \mid} \sum_{\mathcal{S}} \rho^i(\hat{\alpha}, \alpha + \delta_l).$$

3. *Increment consistency*: for the 50 attributes with the highest correlation, we compute

$$\overline{\Delta F_{\delta_k}} = \sum_{i=1}^{50} \sum_{j=1}^{100} \frac{F^i(G(z_j, p_j, \alpha_j + \delta_k)) - F^i(G(z_j, p_j, \alpha_j))}{50 \times 100 \times std(F^i(G(\mathbf{z}, \mathbf{p}, \boldsymbol{\alpha})))},$$

where α_j is the jth original feature vector from a set of 100 samples randomly picked from the validation set, and $\delta_k := \frac{k}{5}, k = 0, 1, ..., 25$. Intuitively, it is defined as the average difference of the predicted attributes of the generated audios (i.e., the difference before and after the attribute increment) as a function of the increment δ_k. We express the result in terms of standard deviations of the non-incremented generated examples as std$(G(z, p, \alpha))$.

*The step of δ_l is defined to obtain more density of points in the range of variation of the attributes (i.e., [0, 1]) as well as $\delta_l > 1$.

4.4.3 Baselines

We compare the metrics described above with *real data* to obtain a baseline for each metric. Also, GANSynth [5], the state-of-the-art on audio synthesis with GANs, is used for comparison. [5] As GANSynth generates 4-second long sounds, the waveform is trimmed down to 1 second for comparison with our models. Additionally, we examine the effect that KD has on these metrics by comparing against a model analogous to DarkGAN, but without using the AudioSet feature conditioning (*baseline*). Experiment results for DarkGAN are shown for different temperature values $T \in \{1, 1.5, 2, 3, 5\}$ (1) as part of the KD process (see Sec. 3.2), and we report separate results for conditional attributes obtained from the training (tr) and validation (val) set.

5. RESULTS

In this section, we present the results from the evaluation procedure described in Sec. 4.4. Furthermore, we validate the quantitative results based on an informal assessment of the generated content.

5.1 Quantitative Metrics

Table 1 presents the metrics scored by DarkGAN$_T$ and the baseline models, as described in Sec. 4.4.3. Note that we condition DarkGAN on attribute vectors randomly sampled from the validation set. Overall, DarkGAN$_{T\in\{1.5,2\}}$ obtains better results than the baselines and is close to *real data* in most metrics. All models score higher PIS than *real data*, with GANSynth in the first place, suggesting that the generated examples have a clear pitch and that the distribution of pitch classes follows that of the training data. This is not surprising, as all the models have explicit pitch conditioning. In contrast, we do not provide conditioning attributes for the instrument class. Therefore, we observe a slight drop in IIS for all models compared to *real data*. DarkGAN$_{T\in\{1.5,2\}}$ achieves the highest IIS, suggesting that the model captured the timbre diversity existent in the dataset and, also, that the generated sounds can be reliably classified into one of all possible instruments. In terms of KID, DarkGAN$_{T\in\{1.5,2\}}$ and *baseline* are on a par with *real data*. A KID equal to *real data* indicates that the Inception embeddings are similarly distributed for real and generated data. As our Inception classifier is trained on pitch and instrument classification and predicting AudioSet features, similarities in such an embedding space indicate common timbral and tonal characteristics between the generated and the real audio data distribution. This trend is maintained in the case of the FAD, where DarkGAN$_{T=2}$ obtains the best scores followed closely by DarkGAN$_{T\in\{1,1.5\}}$.

From the results discussed above, we can conclude that distilling knowledge from the AudioSet classifier helps DarkGAN learning the real data distribution. Furthermore, using slightly higher temperatures in the distillation process yields an improvement over the *baseline* without feature conditioning. We speculate that the additional supervised information that the teacher model provides to DarkGAN's discriminator results in a more meaningful gradient for the generator. Also, attribute conditioning (i.e., attribute vectors sampled from the validation set) may help the generator synthesize diverse samples closer to the training data distribution.

[5] https://github.com/magenta/magenta/tree/master/magenta/models/gansynth

Model	PIS↑		IIS↑		KID↓[a]		FAD↓	
real data	17.7		5.7		6.7		0.1	
GANSynth [5]	**19.6**		4.0		7.1		4.5	
baseline	18.5		4.3		**6.7**		0.8	
DarkGAN$_T$	tr	val	tr	val	tr	val	tr	val
$T=1$	18.4	18.3	4.0	4.0	6.8	6.8	0.7	0.7
$T=1.5$	19.0	19.0	**4.5**	**4.5**	**6.7**	**6.7**	0.7	0.7
$T=2$	19.1	19.0	4.2	4.1	**6.7**	6.8	**0.6**	**0.6**
$T=3$	19.1	19.1	4.2	4.1	6.8	6.8	0.8	0.8
$T=5$	19.2	19.1	4.0	4.0	6.8	6.8	0.8	0.8

[a] $\times 10^{-4}$

Table 1. PIS, IIS, KID and FAD (see Sec. 4.4)

Attribute	T=1	T=1.5	T=2	T=3	T=5
Acoustic guitar	0.20	0.36	**0.39**	0.23	0.10
Bass guitar	0.30	0.38	**0.46**	0.38	0.19
Brass Instrument	0.28	**0.49**	0.38	0.26	0.00
Cello	0.24	**0.29**	0.26	0.17	0.00
Chime	0.15	0.33	**0.39**	0.31	0.03
Guitar	0.28	0.37	**0.42**	0.34	0.13
Plucked string	0.27	0.37	**0.42**	0.32	0.11
Saxophone	0.25	**0.41**	**0.41**	**0.41**	0.03
Trombone	0.18	**0.41**	0.29	0.16	0.00
Trumpet	0.16	**0.46**	0.36	0.25	0.00
...			...		
Didgeridoo	0.06	0.16	**0.21**	0.20	0.08
Drum	0.05	0.21	**0.24**	0.12	0.01
Electronic tuner	0.35	0.44	**0.50**	0.29	0.13
Percussion	0.04	0.19	**0.30**	0.14	0.08
Sine wave	0.28	**0.32**	0.27	0.17	0.10
Singing bowl	0.08	0.20	**0.24**	0.21	0.03
Siren	0.13	0.19	**0.24**	0.10	0.08
Tuning fork	0.22	0.29	**0.35**	0.29	0.10
Zither	0.03	0.18	**0.19**	0.07	-0.01
...			...		
Cat	-0.01	-0.01	-0.01	-0.01	0.00
Chicken, rooster	0.00	-0.06	-0.02	-0.01	-0.01
Domestic animals, pets	-0.01	-0.02	-0.02	0.00	0.00
Frog	0.00	0.03	0.07	0.06	-0.03
Insect	0.00	-0.02	-0.02	-0.02	-0.01
Speech	-0.04	-0.10	-0.07	-0.05	0.01

Table 2. A few examples of attribute correlation coefficients $\rho^i(\hat{\alpha}, \alpha)$ (see Sec. 4.4.2).

5.2 Attribute Consistency and Generalisation

Note that the metrics discussed in this section are not guaranteed to relate directly to human perception, but we consider them suitable indicators of whether the model responds coherently to the input conditioning. There exists the threat of the generator producing adversarial examples, but we argue that this is prevented by the discriminator having to satisfy the Wasserstein criterion (as adversarial examples would exhibit out-of-distribution artifacts). This assumption is also supported by informal listening tests where we find that the metrics correlate with our perception (see Sec. 5.3).

Table 2 shows the results for the *attribute correlation* $\rho^i(\hat{\alpha}, \alpha)$ (see Sec. 4.4.2). At the top of the table, we show a few attributes corresponding to classes represented in the NSynth dataset (e.g., "guitar", "trumpet"). In the middle, we show attributes that, while not being present in the dataset (e.g., "siren", "tuning fork"), still exhibit (relatively) high correlation. At the bottom, attributes that ob-

tain low correlations are presented (e.g., "cat", "insect"). We can observe that models trained with $T \in \{1.5, 2, 3\}$ generally obtain better results than $T \in \{1, 5\}$ in most attributes. Specifically, DarkGAN$_{T=2}$ yields the highest correlations, followed by DarkGAN$_{T=1.5}$. Note that temperatures higher than 1 also improve the correlation for attributes that do not have corresponding classes in the dataset (e.g., "didgeridoo", "percussion", "singing bowl"). This suggests that DarkGAN can extract dark knowledge (which is emphasized by increasing T) from the soft labels. The soft labels indicating the presence of (potentially just slight) timbral characteristics in various sounds are helping the model to learn linearly dependent feature controls for those attributes.

A more in-depth analysis of feature errors and the distribution of features in the dataset would be required to further characterize the results for each attribute. However, it is reasonable that those classes obtaining higher correlations share some timbral features with the training data (e.g., clearly, "violins" are contained in the data set, and a "tuning fork" is similar to a "mallet"). In contrast, those attributes obtaining low correlations may be related to underrepresented features in the training set or features that the model failed to capture.

Fig. 2 shows the correlation coefficient when increasing each attribute by a value δ_l in the input conditioning. The plot reveals that the trend of Table 2 is maintained throughout an ample range of variation of the attributes. Interestingly, while the correlation of DarkGAN$_{T=1}$ considerably declines after an increase $\delta_l > 10^{-0.8}$, using a temperature $T \in \{1.5, 2, 3\}$ the decline is more moderate, and we observe some correlation even for a $\delta_l > 1$, which is outside the range of the attributes.

As the correlation coefficient provides normalized results (regarding scale and offsets), we evaluate the attribute control using the *increment consistency* metric $\overline{\Delta F}_{\delta_k}$ (see Fig. 3). We observe that for low increments of the features ($\delta_k < 1$) temperatures $T \in \{1, 1.5, 2\}$ yield comparable input-output relationships of the features. A temperature $T = 1.5$, however, yields more consistent feature differences for increments $\delta_k > 1$ of the conditional input features. In conclusion, while $DarkGAN_{T=2}$ yields better correlation over all the data (i.e., conditional and predicted attributes are more strongly dependent), for attributes with particularly high correlation, $DarkGAN_{T=1.5}$ performs best in over-emphasizing dark knowledge contained in the data (i.e., the degree of change is higher, especially for $\delta_k > 1$).

5.3 Informal Listening

In the accompanying website,[6] we show sounds generated under various conditioning settings, including generations with feature combinations randomly sampled from the validation set, generations where we fix α and p while changing z, timbre transfer, scales, and more. Overall, we find the results of PIS, IIS, KID, and FAD, discussed in Sec. 5.1, to align well with our perception. The quality of the generated audio is acceptable for all models. Also,

[6] https://an-1673.github.io/DarkGAN.io/

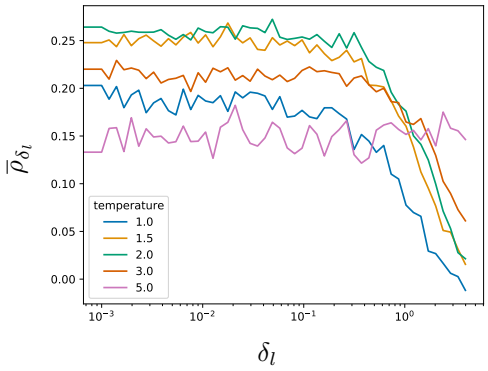

Figure 2. Out-of-distribution average attribute correlation $\overline{\rho}_{\delta_l}$ (see Sec. 4.4.2)

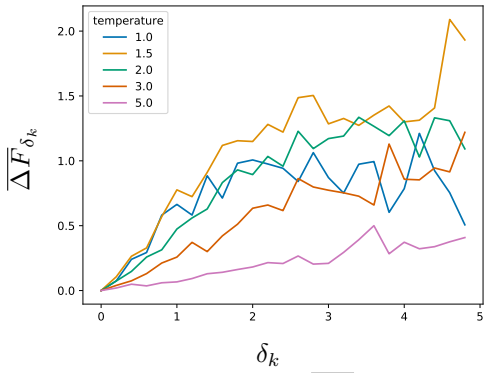

Figure 3. Increment consistency $\overline{\Delta F}_{\delta_k}$ (see Sec.4.4.2)

we find the generated examples to be diverse in terms of timbre, and the tonal content is coherent with the pitch conditioning. Moreover, we perceive that most of the attributes exhibiting high correlations (see Table 2) are audible in the generated output, particularly in the case of DarkGAN$_{T\in\{1,1.5,2\}}$. For higher temperatures $T \in \{3, 5\}$, the model's responsiveness to the attribute conditioning drops substantially. We find the model to be particularly responsive to attributes such as "drum", "tuning fork", "theremin", "choir", or "cowbell". To other attributes (e.g., "accordion", "piano", or "organ"), even though the analysis yields moderate correlations, the model does not seem to produce perceptually satisfactory outputs.

6. CONCLUSION

In this work, we distilled knowledge from a large-scale audio tagging system into DarkGAN, an adversarial synthesizer of tonal sounds. The goal was to enable steering the synthesis process using attributes from the AudioSet ontology. A subset of the NSynth dataset was fed to a pretrained audio tagging system to obtain AudioSet predictions. These predictions were then used to condition DarkGAN. The proposed Knowledge Distillation (KD) framework was evaluated by comparing different temperature settings and employing a diverse set of metrics. Results showed that DarkGAN can generate audio resembling the true dataset and enables moderate control over a comprehensible vocabulary of attributes. By slightly increasing the temperature during the distillation process, we can further improve the responsiveness of the attribute controls. It is also notable that KD can be performed even when the original dataset (i.e., the AudioSet collection) is not involved.

7. ACKNOWLEDGEMENTS

This research is supported by the European Union's Horizon 2020 research and innovation program under the Marie Sklodowska-Curie grant agreement No. 765068 (MIP-Frontiers).

8. REFERENCES

[1] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. of the 28th International Conference on Neural Information Processing Systems NIPS*, Montreal, Quebec, Canada, Dec. 2014.

[2] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of stylegan," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*. Seattle, WA, USA: IEEE, June 2020, pp. 8107–8116.

[3] A. Brock, J. Donahue, and K. Simonyan, "Large scale GAN training for high fidelity natural image synthesis," in *7th International Conference on Learning Representations, ICLR*. New Orleans, LA, USA: OpenReview.net, May 2019.

[4] T. Park, M. Liu, T. Wang, and J. Zhu, "Semantic image synthesis with spatially-adaptive normalization," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*. Long Beach, CA, USA: Computer Vision Foundation / IEEE, June 2019, pp. 2337–2346.

[5] J. Engel, K. K. Agrawal, S. Chen, I. Gulrajani, C. Donahue, and A. Roberts, "Gansynth: Adversarial neural audio synthesis," in *Proc. of the 7th International Conference on Learning Representations, ICLR*, May 2019.

[6] J. Nistal, S. Lattner, and G. Richard, "Drumgan: Synthesis of drum sounds with timbral feature conditioning using generative adversarial networks," in *Proc. of the 21st Int. Society for Music Information Retrieval Conf.*, Montréal, Canada, 2020.

[7] W. S. Peebles, J. Peebles, J. Zhu, A. A. Efros, and A. Torralba, "The hessian penalty: A weak prior for unsupervised disentanglement," in *Computer Vision - ECCV 2020 - 16th European Conference*, ser. Lecture Notes in Computer Science, vol. 12351. Glasgow, UK: Springer, August 2020, pp. 581–597.

[8] A. Voynov and A. Babenko, "Unsupervised discovery of interpretable directions in the GAN latent space," in *Proceedings of the 37th International Conference on Machine Learning, ICML*, ser. Proceedings of Machine Learning Research, vol. 119. Virtual Event: PMLR, July 2020, pp. 9786–9796.

[9] Y. Shen, J. Gu, X. Tang, and B. Zhou, "Interpreting the latent space of gans for semantic face editing," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*. Seattle, WA, USA: IEEE, June 2020, pp. 9240–9249.

[10] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, June 2009.

[11] H. Caesar, J. R. R. Uijlings, and V. Ferrari, "Coco-stuff: Thing and stuff classes in context," in *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR*. Salt Lake City, UT, USA: IEEE Computer Society, June 2018, pp. 1209–1218.

[12] H. Xiao, K. Rasul, and R. Vollgraf, "Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms," *CoRR*, vol. abs/1708.07747, 2017.

[13] J. Engel, C. Resnick, A. Roberts, S. Dieleman, M. Norouzi, D. Eck, and K. Simonyan, "Neural audio synthesis of musical notes with wavenet autoencoders," in *Proc. of the 34th International Conference on Machine Learning, ICML*, Sydney, NSW, Australia, Aug. 2017.

[14] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, "Panns: Large-scale pretrained audio neural networks for audio pattern recognition," *IEEE ACM Trans. Audio Speech Lang. Process.*, vol. 28, pp. 2880–2894, 2020.

[15] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*. New Orleans, LA, USA: IEEE, March 2017, pp. 776–780.

[16] G. E. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *CoRR*, vol. abs/1503.02531, 2015.

[17] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. W. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," in *Proc. of the 9th ISCA Speech Synthesis Workshop*, Sunnyvale, CA, USA, Sept. 2016.

[18] S. Vasquez and M. Lewis, "Melnet: A generative model for audio in the frequency domain," *CoRR*, vol. abs/1906.01083, 2019.

[19] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *Proc. of the 2nd International Conference on Learning Representations, ICLR*, Banff, AB, Canada, Apr. 2014.

[20] C. Aouameur, P. Esling, and G. Hadjeres, "Neural drum machine: An interactive system for real-time synthesis of drum sounds," in *Proc. of the 10th International Conference on Computational Creativity, ICCC*, June 2019.

[21] P. Esling, N. Masuda, A. Bardet, R. Despres, and A. Chemla-Romeu-Santos, "Universal audio synthesizer control with normalizing flows," *Journal of Applied Sciences*, 2019.

[22] P. Esling, A. Chemla-Romeu-Santos, and A. Bitton, "Bridging audio analysis, perception and synthesis with perceptually-regularized variational timbre spaces," in *Proc. of the 19th International Society for Music Information Retrieval Conference, ISMIR*, Paris, France, Sept. 2018.

[23] ——, "Generative timbre spaces with variational audio synthesis," in *Proc. of the 21st International Conference on Digital Audio Effects DAFx-18*, Aveiro, Portugal, Sept. 2018.

[24] P. Dhariwal, H. Jun, C. Payne, J. W. Kim, A. Radford, and I. Sutskever, "Jukebox: A generative model for music," *CoRR*, vol. abs/2005.00341, 2020.

[25] O. Cífka, A. Ozerov, U. Simsekli, and G. Richard, "Self-supervised VQ-VAE for one-shot music style transfer," *CoRR*, vol. abs/2102.05749, 2021.

[26] Y. Saito, S. Takamichi, and H. Saruwatari, "Statistical parametric speech synthesis incorporating generative adversarial networks," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 26, pp. 84–96, 2018.

[27] T. Kaneko and H. Kameoka, "Parallel-data-free voice conversion using cycle-consistent adversarial networks," *CoRR*, vol. abs/1711.11293, 2017.

[28] S. Huang, Q. Li, C. Anil, X. Bao, S. Oore, and R. B. Grosse, "Timbretron: A wavenet(cyclegan(cqt(audio))) pipeline for musical timbre transfer," in *Proc. of the 7th International Conference on Learning Representations, ICLR*, New Orleans, LA, USA, May 2019.

[29] M. Binkowski, J. Donahue, S. Dieleman, A. Clark, E. Elsen, N. Casagrande, L. C. Cobo, and K. Simonyan, "High fidelity speech synthesis with adversarial networks," in *8th International Conference on Learning Representations, ICLR*, Addis Ababa, Ethiopia, April 2020.

[30] J. Kong, J. Kim, and J. Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," in *Annual Conference on Neural Information Processing Systems, NeurIPS*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., Virtual conference, December 2020.

[31] K. Kumar, R. Kumar, T. de Boissiere, L. Gestin, W. Z. Teoh, J. Sotelo, A. de Brébisson, Y. Bengio, and A. C. Courville, "Melgan: Generative adversarial networks for conditional waveform synthesis," in *Proc. of the Annual Conference on Neural Information Processing Systems, NIPS*, Vancouver, BC, Canada, Dec. 2019.

[32] R. Yamamoto, E. Song, and J. Kim, "Parallel wavegan: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*. Barcelona, Spain: IEEE, May 2020, pp. 6199–6203.

[33] C. Donahue, J. McAuley, and M. Puckette, "Adversarial audio synthesis," in *Proc. of the 7th International Conference on Learning Representations, ICLR*, May 2019.

[34] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of gans for improved quality, stability, and variation," in *International Conference on Learning Representations, ICLR*, May 2018.

[35] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of wasserstein gans," in *Proc. of the International Conference on Neural Information Processing Systems, NIPS*, Long Beach, CA, USA, Dec. 2017.

[36] T. Salimans, I. J. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training gans," in *Proc. of the International Conference on Neural Information Processing Systems, NIPS*, Barcelona, Spain, Dec. 2016.

[37] J. Drysdale, M. Tomczak, and J. Hockman, "Adversarial synthesis of drum sounds," in *DAFX*, 2020.

[38] E. Hosseini-Asl, Y. Zhou, C. Xiong, and R. Socher, "A multi-discriminator cyclegan for unsupervised non-parallel speech domain adaptation," in *Proc. of the 19th Annual Conference of the International Speech Communication Association*, Hyderabad, India, Sept. 2018.

[39] D. Michelsanti and Z. Tan, "Conditional generative adversarial networks for speech enhancement and noise-robust speaker verification," in *Proc. of the 18th Annual Conference of the International Speech Communication Association, Interspeech*, Stockholm, Sweden, Aug. 2017.

[40] A. Biswas and D. Jia, "Audio codec enhancement with generative adversarial networks," in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*. Barcelona, Spain: IEEE, May 2020, pp. 356–360.

[41] C. Bucila, R. Caruana, and A. Niculescu-Mizil, "Model compression," in *Proceedings of the Twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, T. Eliassi-Rad, L. H. Ungar, M. Craven, and D. Gunopulos, Eds. Philadelphia, PA, USA: ACM, August 2006, pp. 535–541.

[42] J. Ba and R. Caruana, "Do deep nets really need to be deep?" in *Annual Conference on Neural Information Processing Systems*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, Eds., Montreal, Quebec, Canada, December 2014, pp. 2654–2662.

[43] J. Li, R. Zhao, J. Huang, and Y. Gong, "Learning small-size DNN with output-distribution-based criteria," in *15th Annual Conference of the International Speech Communication Association, INTERSPEECH*, H. Li, H. M. Meng, B. Ma, E. Chng, and L. Xie, Eds. Singapore: ISCA, September 2014, pp. 1910–1914.

[44] N. Papernot, M. Abadi, Ú. Erlingsson, I. J. Goodfellow, and K. Talwar, "Semi-supervised knowledge transfer for deep learning from private training data," in *5th International Conference on Learning Representations, ICLR*, Toulon, France, April 2017.

[45] R. Anil, G. Pereyra, A. Passos, R. Ormándi, G. E. Dahl, and G. E. Hinton, "Large scale distributed neural network training through online distillation," in *6th International Conference on Learning Representations, ICLR*, Vancouver, BC, Canada, May 2018.

[46] M. Yuan and Y. Peng, "CKD: cross-task knowledge distillation for text-to-image synthesis," *IEEE Trans. Multim.*, vol. 22, no. 8, pp. 1955–1968, 2020.

[47] ——, "Text-to-image synthesis via symmetrical distillation networks," in *2018 ACM Multimedia Conference on Multimedia Conference, MM*, S. Boll, K. M. Lee, J. Luo, W. Zhu, H. Byun, C. W. Chen, R. Lienhart, and T. Mei, Eds. Seoul, Republic of Korea: ACM, October 2018, pp. 1407–1415.

[48] W. Chan, N. R. Ke, and I. Lane, "Transferring knowledge from a RNN to a DNN," in *16th Annual Conference of the International Speech Communication Association, INTERSPEECH*. Dresden, Germany: ISCA, September 2015, pp. 3264–3268.

[49] Z. Tang, D. Wang, and Z. Zhang, "Recurrent neural network training with dark knowledge transfer," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*. Shanghai, China: IEEE, March 2016, pp. 5900–5904.

[50] T. Asami, R. Masumura, Y. Yamaguchi, H. Masataki, and Y. Aono, "Domain adaptation of DNN acoustic models using knowledge distillation," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*. New Orleans, LA, USA: IEEE, March 2017, pp. 5185–5189.

[51] Y. Aytar, C. Vondrick, and A. Torralba, "Soundnet: Learning sound representations from unlabeled video," in *Annual Conference on Neural Information Processing Systems, NeurIPS*, Barcelona, Spain, December 2016, pp. 892–900.

[52] L. Gao, K. Xu, H. Wang, and Y. Peng, "Multi-representation knowledge distillation for audio classification," *CoRR*, vol. abs/2002.09607, 2020.

[53] G. Hinton, O. Vinyals, and J. Dean, "Dark knowledge," in *Toyota Technological Institute at Chicago, TTIC*, 2014.

[54] J. Nistal, S. Lattner, and G. Richard, "Comparing representations for audio synthesis using generative adversarial networks," in *Proc. of the 28th European Signal Processing Conference, EUSIPCO2020*, Amsterdam, NL, Jan. 2021.

[55] A. Odena, C. Olah, and J. Shlens, "Conditional image synthesis with auxiliary classifier gans," in *Proc. of the 34th International Conference on Machine Learning, ICML*, Sydney, NSW, Australia, Aug. 2017.

[56] C. Gupta, P. Kamath, and L. Wyse, "Signal representations for synthesizing audio textures with generative adversarial networks," *CoRR*, vol. abs/2103.07390, 2021.

[57] M. Binkowski, D. J. Sutherland, M. Arbel, and A. Gretton, "Demystifying MMD gans," in *Proc. of the 6th International Conference on Learning Representations, ICLR*, Vancouver, BC, Canada, Apr. 2018.

[58] K. Kilgour, M. Zuluaga, D. Roblek, and M. Sharifi, "Fréchet audio distance: A metric for evaluating music enhancement algorithms," *CoRR*, vol. abs/1812.08466, 2018.