Face-to-Music Translation

Chelhwon Kim* Leia Inc.

Andrew Port* University of California, Santa Cruz

Mitesh Patel NVIDIA

Abstract

Learning a mapping between two unrelated domains-such as image and audio, without any supervision is a challenging task. In this work, we propose to use a distance-preserving generative adversarial model to translate images of human faces into an audio domain. The audio domain is defined by a collection of musical note sounds recorded by 10 different instrument families (NSynth [5]) and a distance metric where the instrument family class information is incorporated together with a mel-frequency cepstral coefficients (MFCCs) feature. To enforce distance-preservation, a loss term that penalizes the difference between pairwise distances of the faces and the translated audio samples is used. Further, we discover that the distance preservation constraint in the generative adversarial model leads to reduced diversity in the translated audio samples, and propose the use of an auxiliary discriminator to enhance the diversity of the translations while using the distance preservation constraint. We also provide a visual demonstration of the results (video demo) and numerical analysis of the fidelity of the translations.

1. Introduction

There has been a lot of work in attempting to find a meaningful mapping between two different domains without any supervision - i.e. only unordered and unpaired samples from the two domains are given. However, most of them are often limited to a task of mapping between two visual domains such as image-to-image translation [6, 9, 20], where the learner takes an image in one domain and maps it into another pixel domain with its content or style changed to be similar to that of the target samples. In this work, we focus on finding a mapping between two highly unrelated domains - image and audio. More precisely, we want to find a meaningful mapping that links images of human faces into musical sounds, while the measure of similarity between face images is consistent with the one between translated audio samples so that visually dis/similar face images can be mapped into audio samples that sound perceptually dis/similar to each other. This could be used to aid those with a visual impairment by encoding the visual information (e.g. facial appearance) into the audio domain, which allows them to perceive this information using their ears [15, 11].

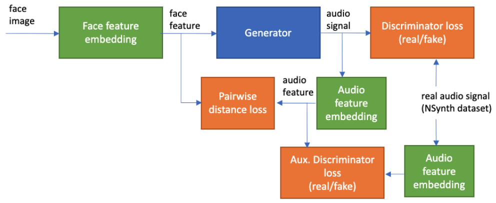


Figure 1. Our proposed image-to-audio translation model's structure.

To find a meaningful mapping between such unrelated domains, we adapt the distance-preserving technique [1] to map images of faces into an audio domain defined by a collection of musical note sounds recorded by 10 different instrument families (NSynth [5]), and a new audio metric we designed, where the instrument family class information is incorporated together with the MFCCs, as a result, musical samples of the same instrument family with similar timbre have small distances and vice versa. Further, our model also demonstrates that, when translating between such unrelated modalities, there is a trade-off between the variety of the translations (i.e. audio), and the preservation of geometric information. To address this problem, we propose using an auxiliary discriminator in a way that it further enforces that the model outputs sounds which fit into the target audio dataset in the new designed audio metric space.

2. Method

Our method employs the distance-preserving mapping technique [1] in the GANs framework: We first find a feature embedding model that embeds faces into a metric space, where the distance corresponds to a measure of face similarity. This embedding model can be obtained by refining a convolutional neural network that is pretrained for the face recognition task on a large-scale face dataset (VGGFace2 [2]), where the network captures information about the facial appearance to successfully classify the face identities. We then train a deep generative model by using a GAN approach [7] that takes the face features from that embedding space [1] and synthesizes raw audio waveforms that fit into the distribution of sounds specified by a given target dataset (i.e. musical notes of NSynth dataset [5]). Two dis-

*equal contribution

[1] Typically GANs take as input a latent vector sampled from a Gaussian or uniform distribution. For our task, we are using the pretrained feature embedding space as the input latent.

criminators are simultaneously trained with the generator, whose task is to predict whether the generated output is real or fake. See the overview of our model's structure in Fig. 1.

Metric Preservation Constraint: To enforce the distance preservation constraint, we add a metric loss term (Eq. 1) to the GAN's adversarial loss, which computes the difference of pairwise distances of the face features and features of the generated audio samples ('Pairwise distance loss' block in Fig. 1).

$$\mathcal{L}_{metric} = \frac{1}{Z}\sum_{i<j}\left|\frac{||f(x_i)-f(x_j)||_2-\mu_x}{\sigma_x}-\frac{||\phi(y_i)-\phi(y_j)||_2-\mu_y}{\sigma_y}\right| \tag{1}$$

where Z is the number of possible (unordered) pairs of samples in the training mini-batch, $f(x)$ is the face embedding feature of the input face image x, $\phi(y)$ denotes our audio embedding feature of the translated audio output y. The standardization parameters, (μ_x, σ_x) and (μ_y, σ_y), are derived from source (VGGFace2) and target (NSynth) datasets respectively and are the mean and standard deviation of pairwise distances of samples from within each respective dataset.

Audio Metric: The audio feature is computed by a combination of a pretrained audio feature embedding model and MFCCs (The 'Audio feature embedding' block in Fig. 1): We first train a feature embedding model for audio that maps raw audio waveforms into a compact Euclidean space where the distance directly corresponds to a measure of audio similarity. This is done by using the triplet loss function [18] that aims to separate the positive sample from the negative sample by a distance margin, where the positive and negative are determined by the annotated instrument family class label of the NSynth dataset. As a result, audio samples that have the same musical instrument timbre are mapped into features with small distances and vice versa. To further incorporate perceptual notions of distance into the audio metric, the Mel-frequency cepstral coefficients (MFCCs) [19] of the audio are computed and concatenated with the learned audio features. Fig. 2 (b) shows the audio metric space of NSyth samples. For the visualization, we mapped them into a 2D space by tSNE [10]. Each blob represents an NSynth sample's feature vector and is color-coded by the instrument family (bass, brass, flute, guitar, keyboard, mallet, organ, reed, string, and vocal).

Enhancement of the Sample Variety: The adversarial loss enforces that the generated audio fits into the given target audio dataset. Despite this, when using a metric preservation constraint, we observe decreased variety in the generated audio. We verify this perceived lack of variety using Frêchet Inception Distance (FID) [8] [2]. Our FID score results and a visual check can be seen in Table 1 and Figure 2 respectively. To address this, we add an auxiliary discriminator as an additional adversarial criterion that can support the main discriminator in a way that it further encourages the generative model to output audio samples that fit into the distribution of the target audio dataset. Note that, unlike the main discriminator, we attach this second discriminator on top of the hybrid MFCC-learned audio embedding feature (Fig. 1 'Audio feature embedding' block) to predict the real/fake by processing the audio features rather than the raw audio signals. As we will see in the next section, this helps the generative model output audio samples that are spread over the distribution of real audio samples in our audio metric space (Fig. 2 (b) third row). Furthermore, our additional discriminator has a much simpler and more economical architecture than the main discriminator, consisting of five fully connected layers with dimension 128, 64, 32, 16, and 1.

Table 1. The quality measure scores of Frêchet Inception Distance (FID, lower is better) and Inception Score (IS, higher is better) by a pitch classifier and a family classifier, the distance preservation measure score by Pearson Correlation (PC, higher is better), and the clustering measure score by Silhouette Value (SV, higher is better).

	Pitch		Family			
Method	**FID**	**IS**	**FID**	**IS**	**PC**	**SV**
Baseline	21.84	56.27	51.02	4.59	0.090	-0.0051
+metric preservation	46.69	53.18	90.99	3.98	0.683	0.0112
+aux. discriminator	25.52	54.97	53.98	5.89	0.432	0.0256

3. Results

In this section, we detail an ablation study performed on our proposed model by adding sequentially the metric preservation loss and then the second discriminator to the baseline to see the effect of each constraint. See Table 1.

The baseline model is based on WaveGAN [3]. It is trained to, given a 512-dimensional VGGFace2 feature vector, generate a raw audio signal of length 8192 that fits into the distribution of NSynth training samples. To improve the quality of the generated sounds, we employ a state-of-the-art stabilization technique, spectral normalization [12]. Further, our baseline model is conditioned by a 'pitch' label given by the NSynth dataset (i.e. our model is a *conditional* GAN similar to that described in [4]) but with conditional batch normalization layers [14] in the generative network and a projection layer at the end of the discriminator as in [13].

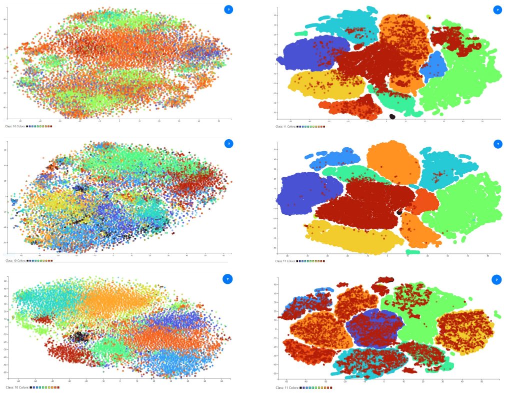
Table 1 shows four evaluation metric scores: 1) Pearson product momentum correlation (PC) between L2 pairwise distances of the samples in the source metric and the corresponding translated samples in the audio metric. The Pearson correlation has a value between +1 and -1, where

[2] FID is known to be sensitive to the variety of the generated samples.

1 is perfect positive linear correlation. 2) Silhouette value (SV) to inversely check whether the clusters are preserved in the source metric space. More precisely, we first assign the class label to each translated audio clip based on its closest cluster of “real” audio samples in the audio metric space. I.e. a translated audio sample which lies within or close to the ‘string’ cluster will be assigned the ‘string’ family label. Then, we measure how well the corresponding untranslated samples are grouped by their assigned class labels in the source metric space (Fig. 2 (a) Face metric space). To measure this quantitatively, we use Silhouette value. A method which was proposed by [16] and has been used as a means for clustering evaluation. 3) The well known Inception Score (IS) [17] and Frêchet Inception Distance (FID) [8] to measure the quality of the translated audio samples.

First, the baseline model achieves the FID and IS scores of 21.84 (51.02) and 56.27 (4.59) respectively by our pretrained pitch (family) classification model [3]. With the distance metric preservation constraint on the model, the Pearson Correlation (PC) score increases from 0.090 to 0.683. This demonstrates that our metric preservation loss helps the model preserve the geometric structure of the face embeddings in the target audio metric space, however, we also observed a reduced quality in the samples, i.e FID increases from 21.84 (51.02) to 46.69 (90.00), and IS decreases from 56.27 (4.59) to 53.18 (3.98). Our visualization of the translated audio samples in the audio metric space also demonstrates this. See Fig. 2 (b). The red dots represent the translated audio clips from a subset of 20k randomly sampled face images from the VGGFace2 train set and the blobs are the NSynth real audio samples. The first two top figures correspond to the baseline and the our proposed model with metric preservation loss. It is apparent that the distribution of red dots (i.e. translated audio samples) shrinks and forms a cluster surrounded by other “real” clusters and we find that most of those generated audio samples play interpolated sounds between the nearby instrument families (clusters) and hard to find samples that play other instrument families at a distance from them (e.g. the right most green cluster (‘mallet’) and left top light-blue cluster (‘flute’)).

By adding an additional adversarial constraint by the auxiliary discriminative network, we improve the variety while it still preserves the source metric: 25.52 (FID) and 0.432 (PC). We also observed that the auxiliary discriminator changes a lot the translations’ distribution in the audio metric space in a way that they are spread evenly over the real clusters (See the third plot in Fig. 2 (b)) and the translations play much wider variety of musical sounds (so FID score decreased).



(a) Face metric space (b) Audio metric space

Figure 2. (a) tSNE visualization of the source face embedding metric space with 20k VGGFace2 face samples. Each face sample is color-coded by its estimated instrument family label (See text for details). (b) tSNE visualization of the target audio metric space with the NSynth real audio samples (color-coded blobs) and translated face samples (red dots). From Top to Bottom: 1) baseline model 2) w/ metric preservation 3) w/ auxiliary discriminator. The colors represent the 10 instrument families: bass, brass, flute, guitar, keyboard, mallet, organ, reed, string, and vocal.

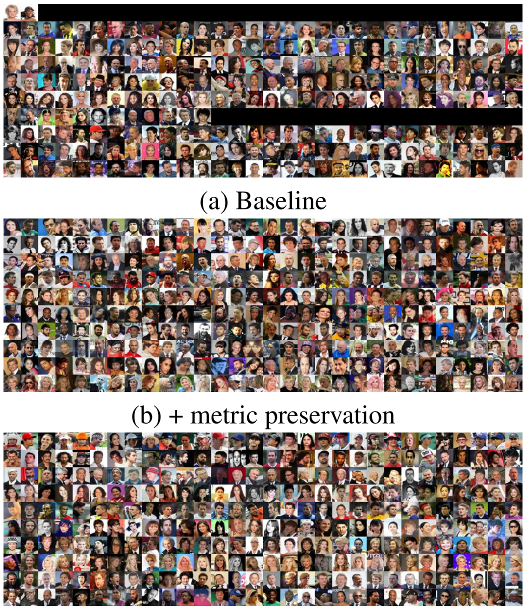
We also measured that how well faces are grouped by their assigned instrument class labels in the source metric space by the Silhouette value and our model with the auxiliary discriminator outperforms others by 0.0256. Note that overall SV scores are low since the pre-trained face embedding was not designed to have such well separated clusters, but the SV rewards the result where faces with the same class sit close together in the space. Fig. 2 (a) shows the same randomly sampled 20k face features in the source metric space color-coded by the assigned labels. Our translation model with the auxiliary discriminator (the third plot in Fig. 2 (a)) clearly shows the clusters by the assigned labels.

We also randomly sampled face images belonging to the instrument family label to visually check if those faces show a pattern on their visual appearance (e.g. same skin tone, hair color etc.). See Fig. 3: each row shows face images belonging to the same instrument family (i.e. from the dots with the same color in Fig. 2 (a)). Our translation model (Fig. 3 (c)) shows faces belonging to the same instrument family are visually similar to each other and show one or more common attributes (e.g. female/blond hair in the eighth row).

4. Conclusions

We have proposed a distance preserving generative adversarial network that automatically learns an information preserving embedding between two unrelated domains of media (i.e. image and audio domain). We discovered that there is a trade-off between the variety of translations and the preservation of geometric information. To address this

[3] Note that the maximum of IS score is the number of classes which is 61 for the pitch class and 10 for the family class. The minimum score of FID is zero for the both classification models.



(a) Baseline

(b) + metric preservation

(c) + auxiliary discriminator

Figure 3. Randomly sampled face images grouped by their instrument class, i.e. a visual check to show that faces assigned to the same instrument class share visual attributes (e.g. same skin tone, hair color, etc.). Our translation model (c) shows faces belonging to the same instrument family are visually similar to each other and show one or more common attributes (e.g. female/blond hair in the seventh row). If the number of faces per class is fewer than 30, that row is padded by black pixels.

problem, we proposed to use the auxiliary discriminator that can support the primary discriminator, which helps to balance the performance between the diversity and the metric preservation. We demonstrate that the proposed model translates a face image from the VGGFace2 dataset into a musical sound that plays one of 10 instrument family in the NSynth dataset and the faces playing the same instrument type show a pattern on their visual appearance (e.g skin tone, hair color etc.).

References

[1] S. Benaim and L. Wolf. One-sided unsupervised domain mapping. In *Advances in neural information processing systems*, pages 752–762, 2017. 1

[2] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *International Conference on Automatic Face and Gesture Recognition*, 2018. 1

[3] C. Donahue, J. McAuley, and M. Puckette. Adversarial audio synthesis. *arXiv preprint arXiv:1802.04208*, 2018. 2

[4] J. Engel, K. K. Agrawal, S. Chen, I. Gulrajani, C. Donahue, and A. Roberts. Gansynth: Adversarial neural audio synthesis. *arXiv preprint arXiv:1902.08710*, 2019. 2

[5] J. Engel, C. Resnick, A. Roberts, S. Dieleman, D. Eck, K. Simonyan, and M. Norouzi. Neural audio synthesis of musical notes with wavenet autoencoders, 2017. 1

[6] L. A. Gatys, A. S. Ecker, and M. Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2414–2423, 2016. 1

[7] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014. 1

[8] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in neural information processing systems*, pages 6626–6637, 2017. 2, 3

[9] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017. 1

[10] L. v. d. Maaten and G. Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008. 2

[11] P. Meijer. An experimental system for auditory image representations. *Biomedical Engineering, IEEE Transactions on*, 39:112 – 121, 03 1992. 1

[12] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida. Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957*, 2018. 2

[13] T. Miyato and M. Koyama. cgans with projection discriminator. In *International Conference on Learning Representations*, 2018. 2

[14] E. Perez, F. Strub, H. De Vries, V. Dumoulin, and A. Courville. Film: Visual reasoning with a general conditioning layer. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018. 2

[15] A. Port, C. Kim, and M. Patel. Earballs: Neural transmodal translation. *arXiv preprint arXiv:2005.13291*, 2020. 1

[16] P. J. Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987. 3

[17] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen. Improved techniques for training gans. In *Advances in neural information processing systems*, pages 2234–2242, 2016. 3

[18] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015. 2

[19] S. S. Stevens, J. Volkmann, and E. B. Newman. A scale for the measurement of the psychological magnitude pitch. *The Journal of the Acoustical Society of America*, 8(3):185–190, 1937. 2

[20] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, 2017. 1