Play as You Like: Timbre-Enhanced Multi-Modal Music Style Transfer

Chien-Yu Lu,[1]* Min-Xin Xue,[1]* Chia-Che Chang,[1] Che-Rung Lee,[1] Li Su[2]
[1]Department of Computer Science, National Tsing-Hua University, Hsinchu, Taiwan
[2]Institute of Information Science, Academia Sinica, Taipei, Taiwan
{j19550713, liedownisok, chang810249}@gmail.com, cherung@cs.nthu.edu.tw, lisu@iis.sinica.edu.tw

Abstract

Style transfer of polyphonic music recordings is a challenging task when considering the modeling of diverse, imaginative, and reasonable music pieces in the style different from their original one. To achieve this, learning stable multi-modal representations for both domain-variant (i.e., style) and domain-invariant (i.e., content) information of music in an unsupervised manner is critical. In this paper, we propose an unsupervised music style transfer method without the need for parallel data. Besides, to characterize the multi-modal distribution of music pieces, we employ the Multi-modal Unsupervised Image-to-Image Translation (MUNIT) framework in the proposed system. This allows one to generate diverse outputs from the learned latent distributions representing contents and styles. Moreover, to better capture the granularity of sound, such as the perceptual dimensions of timbre and the nuance in instrument-specific performance, cognitively plausible features including mel-frequency cepstral coefficients (MFCC), spectral difference, and spectral envelope, are combined with the widely-used mel-spectrogram into a timbre-enhanced multi-channel input representation. The Relativistic average Generative Adversarial Networks (RaGAN) is also utilized to achieve fast convergence and high stability. We conduct experiments on bilateral style transfer tasks among three different genres, namely piano solo, guitar solo, and string quartet. Results demonstrate the advantages of the proposed method in music style transfer with improved sound quality and in allowing users to manipulate the output.

Introduction

The music style transfer problem has been receiving increasing attention in the past decade (Dai and Xia 2018). When discussing this problem, typically we assume that music can be decomposed into two of its attributes, namely *content* and *style*, the former being domain-invariant and the latter domain-variant. This problem is therefore considered as to modify the style of a music piece while preserving its content. However, the boundary that distinguishing content and style is highly dynamic; different objective functions in timbre, performance style or composition are related to different style transfer problems (Dai and Xia 2018). Traditional style transfer methods based on feature interpolation (Caetano and Rodet 2011) or matrix factorization (Driedger, Prätzlich, and Müller 2015; Su et al. 2017) typically need a *parallel* dataset containing musical notes in the target-domain style, and every note has a pair in the source domain. In other words, we need to specify the content attribute element-wisely, and make style transfer be performed in a supervised manner. Such restriction highly limits the scope that the system can be applied. To achieve higher-level mapping across domains, recent approaches using deep learning methods such as the generative adversarial networks (GAN) (Goodfellow et al. 2014) allow a system to learn the content and style attributes directly from data in an unsupervised manner with extra flexibility in mining the attributes relevant to content or style (Ulyanov and Lebedev 2016; Bohan 2017; Wu et al. 2018; Verma and Smith 2018; Haque, Guo, and Verma 2018; Mor et al. 2018).

Beyond the problem of unsupervised domain adaptation, there are still technical barriers concerning realistic music style transfer applicable for various kinds of music. First, previous studies can still hardly achieve multi-modal and non-deterministic mapping between different domains. However, when we transfer a piano solo piece into guitar solo, we often expect the outcome of the guitar solo to be *adjustable*, perhaps with various fingering styles, brightness, musical texture, or other sound quality. Second, the transferred music inevitably undergoes degradation of perceptual quality such as severely distorted musical timbre; this indicates the need of a better representation for timbre information. Although many acoustic correlates of timbre have been verified via psychoacoustic experiments (Grey 1977; Alluri and Toiviainen 2010; Caclin et al. 2005) and also been used in music information retrieval (Lartillot, Toiviainen, and Eerola 2008; Peeters et al. 2011), they are rarely discussed in deep-learning-based music style transfer problems. This might be because of several reasons: some acoustic correlates are incompatible to the format of modern deep learning architectures; rawer data inputs such as waveforms and spectrograms are still preferred to reveal the strength of deep learning; and even, an exact theory of those acoustic correlates on human perception is still not clear in cognitive science (Siedenburg, Fujinaga, and McAdams 2016; Aucouturier and Bigand 2013). For this issue, a recently proposed method in (Mor et al. 2018) adopts the WaveNet (Van

*The first two authors are with equal contribution.

Den Oord et al. 2016), the state-of-the-art waveform generator on raw waveform data to generate realistic outputs for various kinds of music with a deterministic style mapping, at the expense of massive computing power.

To address these issues, we consider the music style transfer problem as learning a multi-modal conditional distribution of style in the target domain given only one unpaired sample in the source domain. This is similar to the Multimodal Unsupervised Image-to-Image Translation (MUNIT) problem, where a principled framework proposed in (Huang et al. 2018) is employed in our system. During training, cognitively plausible timbre features including mel-frequency cepstral coefficients (MFCC), spectral difference, and spectral envelope, all designed to have the same dimension with mel-spectrogram, are combined together into a multi-channel input representation in the timbre space. Since these features have close-form relationship with each other, we introduce a new loss function, named *intrinsic consistency loss*, to keep the consistency among the channel-wise features in the target domain. Experiments show that with such extra conditioning on the timbre space, the system does achieve better performance in terms of content preservation and sound quality than those using only the spectrogram. Moreover, comparing to other style transfer methods, the proposed multi-modal method can stably generate diverse and realistic outputs withs improved quality. Also, in the learned representations, some dimensions that disentangle timbre can be observed. Our contributions are two-fold:

- We propose an unsupervised multi-modal music style transfer system for one-to-many generation. To the best of our knowledge, this have not been done before in music style transfer. The proposed system further allows music style transfer from scratch, without massive training data.
- We design multi-channel timbre features with the proposed intrinsic consistency loss to improve the sound quality for better listening experience of the style-transferred music. Disentanglement of timbre characteristics in the encoded latent space is also observed.

Related Works

Generative Adversarial Networks

Since its invention in (Goodfellow et al. 2014), the GAN has shown amazing results in multimedia content generation in variant domains (Yu et al. 2017; Gwak et al. 2017; Li et al. 2017). A GAN comprises two core components, namely the generator and the discriminator. The task of the generator is to fool the discriminator, which distinguishes real samples from generated sample. This loss function, named *adversarial loss*, is therefore implicit and is defined only by the data. Such a property is particularly powerful for generation tasks.

Domain Adaptation

Recent years has witnessed considerable success in unsupervised domain adaptation problems without parallel data, such as image colorization (Larsson, Maire, and Shakhnarovich 2016; Zhang, Isola, and Efros 2016) and image enhancement (Chen et al. 2018). Two of the most popular methods that achieve unpaired domain adaptation could be the CycleGAN (Zhu et al. 2017a) and the Unsupervised Image-to-Image Translation Networks (UNIT) (Liu, Breuel, and Kautz 2017) framework, the former introduce the cycle consistency loss to train with unpaired data and the other is to learn a joint distribution of images in different domains. However, most of these transfer models are based on a deterministic or one-to-one mapping. Therefore, these models are unable to generate diverse outputs when given the data from source domain. One of the earliest attempts on multi-modal unsupervised translation could be (Zhu et al. 2017b), which aims at capturing the distribution of all possible outputs, that means, a one-to-many mapping that maps a single input into multiple outputs. To handle multi-modal translation, two possible methods are: adding random noise to the generator, or adding dropout layer into the generator for capturing the distribution of outputs. However, these methods still tend to generate similar outputs since the generator is easy to ignoring random noise and additional dropout layers. In this paper, we use a disentangled representation framework, MUNIT (Huang et al. 2018), for generating high-quality and high-diversity music pieces with unpaired training data.

Music Style Transfer

The music style transfer problem has been investigated for decades. Broadly speaking, the music being transferred can be either audio signals or symbolic scores (Dai and Xia 2018). In this paper, we focus on the music style transfer of audio signals, where its domain-invariant *content* typically refer to the structure established by the composer (e.g., mode, pitch, or dissonance)[1], and its domain-variant *style* refers to the interpretation of the performer (e.g., timbre, playing styles, expression).

With such abundant implications of content and style, the music style transfer problem encompasses extensive application scenarios, including audio mosaicking (Driedger, Prätzlich, and Müller 2015), audio antiquing (Välimäki et al. 2008; Su et al. 2017), and singing voice conversion (Kobayashi et al. 2014; Wu et al. 2018), to name but a few. Recently, motivated by the success of image style transfer (Gatys, Ecker, and Bethge 2016), using deep learning for music or speech style transfer on audio signals has caught wide attention. These solutions can be roughly categorized into two classes. The first class takes spectrogram as input and feeds it into convolutional neural networks (CNN), recurrent neural networks (RNN), GAN or autoencoder (Haque, Guo, and Verma 2018; Donahue, McAuley, and Puckette 2018). Cycle consistency loss has also been applied for such features (Wu et al. 2018; Hosseini-Asl et al. 2018). The second class takes raw waveform as input and feed it into autoregressive models such as WaveNet (Mor et al. 2018). Unlike the classical approaches, the deep learning

[1]Although the instrumentation process is usually done by the composer, especially in Western classical music, we presume that the timbre (i.e., the instrument chosen for performance) is determined by the performer.

approaches pay less attention to the level of signal processing, and tends to overlook timbre-related features that are psychoacoustically meaningful in describing music styles. One notable exception is (Verma and Smith 2018), which took the deviation of temporal and frequency energy envelopes respectively from the style audio into the loss function of the network, and demonstrated promising results.

Data Representation

We discuss the audio features before introducing the whole framework of the proposed system. We set two criteria of choosing features for our system input. First, all the features can be of the same dimension, so as to facilitate a CNN-based multi-channel architecture, where one feature occupy one input channel. In other words, the channel-wise features represent the colors of sound; this is similar to the case of image processing, where three colors (i.e., R, G, and B) are also taken as channel-wise input. Second, the chosen features should be related to music perception or music signal synthesis. The features verified to be highly correlated to one or more attributes of musical timbre through perceptual experiments are preferred more. As a result, we consider the following four data representations: 1) mel-spectrogram, 2) mel-frequency cepstral coefficients (MFCC), 3) spectral difference, and 4) spectral envelope.

Consider an input signal $\mathbf{x} := \mathbf{x}[n]$ where n is the index of time. Give a N-point window function $\mathbf{h}$ for the computation of the short-time Fourier transform (STFT):

$$\mathbf{X}[k,n] := \sum_{m=0}^{N-1} \mathbf{x}[m+nH]\mathbf{h}[m]e^{-\frac{j2\pi km}{N}} . \quad (1)$$

where k is the frequency index. The sampling rate is $f_s = 22.05$ kHz. We consider the *power spectrogram* of $\mathbf{x}$ being the γ-power of the magnitude part of the STFT, namely $|\mathbf{X}|^\gamma$. In this paper we set $\gamma = 0.6$, a value that well approximate the perceptual scale based on the Stevens power law (Stevens 1957). The mel-spectrogram $\tilde{\mathbf{X}}[f,n] := \mathbf{M}|\mathbf{X}|^\gamma$ is the power spectrogram mapped into the mel-frequency scale with a filterbank. The filterbank $\mathbf{M}$ has 256 overlapped triangular filters ranging from zero to 11.025 kHz, and the filters are equally-spaced in the mel scale: mel $:= 2595\log_{10}(f/700+1)$. MFCC is represented as the discrete cosine transform (DCT) of the mel-spectrum:

$$\mathbf{C}[q,n] := \sum_{f=0}^{F-1} \tilde{\mathbf{X}}[f,n] \cos\left[\frac{\pi}{N}\left(f+\frac{1}{2}\right)q\right] . \quad (2)$$

where q is the cepstral index and $F = 256$ is the number of frequency bands. The MFCC has been one of the most widely used audio feature ranging from a wide diversity of tasks including speech recognition, speaker identification, music classification, and many others. Traditionally, only the first few coefficients of the MFCC are used, as these coefficients are found relevant to timbre-related information. High-quefrency coefficients are then related to pitch. In this work, we adopt all coefficients for end-to-end training.

The *spectral difference* is a classic feature for musical onset detection and timbre classification. It is highly relevant to the attack in the attack-decay-sustain-release (ADSR) envelope of a note. The spectral difference is represented as

$$\Delta\tilde{\mathbf{X}}[f,n] := \mathrm{ReLU}(\tilde{\mathbf{X}}[f,n+1] - \tilde{\mathbf{X}}[f,n]) \quad (3)$$

where ReLU refers to a rectified linear unit that discards the energy-decreasing parts in the time-frequency plane. The accumulation of spectral difference over the frequency axis is the well-known *spectral flux* for musical onset detection.

The *spectral envelope* $\mathbf{Y}$ can be loosely estimated through the inverse DCT of the first η elements of the MFCC, which represents the slow-varying counterpart in the spectrum:

$$\mathbf{Y}[f,n] := \sum_{q=0}^{\eta} \mathbf{C}[q,n] \cos\left[\frac{\pi}{N}\left(q+\frac{1}{2}\right)f\right] , \quad (4)$$

where η is the *cutoff cepstral index*. In this paper we set $\eta = 15$. The spectral envelope has been a well-known factor in timbre and is widely used in sound synthesis []. These data representations emphasize different aspects of timbre, and at the same time able to act as a channel for joint learning.

Proposed Method

Consider the style transfer problem from two domains $\mathcal{X}$ and $\mathcal{Y}$. $x \in \mathcal{X}$ and $y \in \mathcal{Y}$ are two samples from $\mathcal{X}$ and $\mathcal{Y}$, respectively. Assume that the latent spaces of the two domains are partially shared: each x is generated by a content code $c \in \mathcal{C}$ shared by both domains and a style code $s \in \mathcal{S}$ in the individual domain. Inferring the marginal distributions of c and s, namely $p(c)$ and $p(s)$, respectively, therefore allows one to achieve *one-to-many* mapping between $\mathcal{X}$ and $\mathcal{Y}$. This idea was first proposed in the MUNIT framework (Huang et al. 2018). To further improve its performance and to adapt to our problem formulation, we make two extensions. First, to stabilize the generation result and speed up the convergence rate, we adopt the Relativistic average GAN (RaGAN) (Jolicoeur-Martineau 2018) instead of the for the conventional GAN component for generation. Second, considering the relation between the channel-wise timbre features, we introduce the *intrinsic consistency loss* to pertain the relation between the output features.

Overview

Fig. 1 conceptually illustrates the whole multi-mdoal music style transfer architecture. It contains encoders E and generators G for domains $\mathcal{X}$ and $\mathcal{Y}$, namely $E_\mathcal{X}$, $E_\mathcal{Y}$, $G_\mathcal{X}$, and $G_\mathcal{Y}$.[2] E encodes a music piece into a style code s and a content code c. G decodes c and s into the transferred result, where c and G are from different domains and s in the target domain is sampled from a Gaussian distribution $z \in \mathcal{N}(0,1)$. For example, the process $v = G_\mathcal{Y}(c_x, s_y)$ where $s_y \in \mathcal{N}(0,1)$ transfer x in domain $\mathcal{X}$ to v in domain $\mathcal{Y}$. Similarly, the process transferring y in domain $\mathcal{Y}$ to u in domain $\mathcal{X}$ is also shown in Fig. 1.

The system has two main networks, cross-domain translation and within-domain reconstruction, as shown in the left

[2] Since the transfer task is bilateral, we will ignore the subscript if we do not specifically mention $\mathcal{X}$ or $\mathcal{Y}$ domains. For example, G refers to either $G_\mathcal{X}$ or $G_\mathcal{Y}$

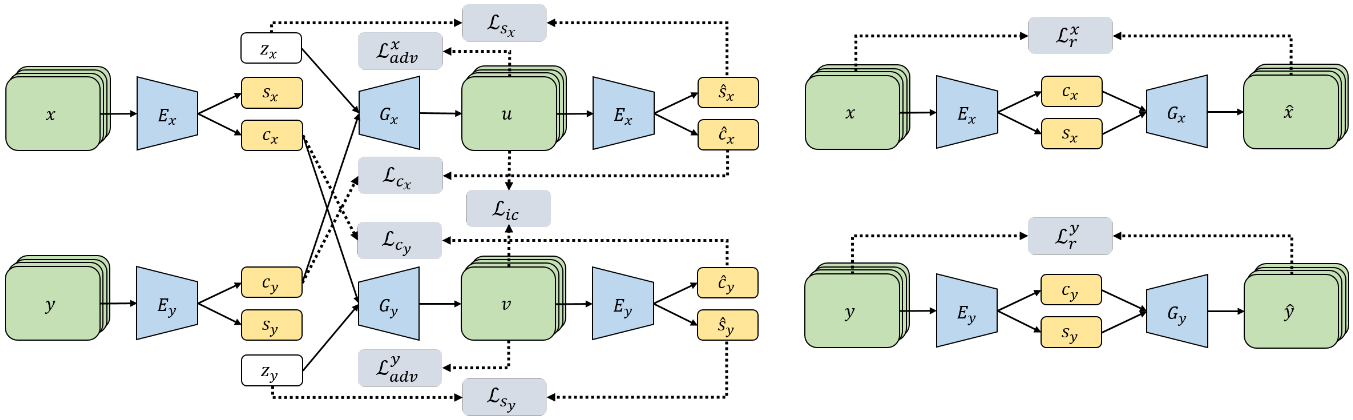

Figure 1: The proposed multi-modal music style transfer system with intrinsic consistency regularization $\mathcal{L}_{ic}$. Left: cross-domain architecture. Right: self-reconstruction architecture.

and the right of Fig. 1, respectively. The cross-domain translation network uses GANs to match the distribution of the transferred features to the distribution of the features in the target domain. It means, discriminators D should distinguish the transferred samples from the ones truly in the target domain, and G needs to fool D by capturing the distribution of the target domain.

By adopting the Chi-Square loss (Mao et al. 2017) in the GANs, the resulting adversarial loss, $\mathcal{L}_{adv}$, is represented as:

$$\begin{aligned}\mathcal{L}_{adv} &= \mathcal{L}_{adv}^x + \mathcal{L}_{adv}^y \\ &= \mathbb{E}_{c_y \sim p(c_y), z \sim \mathcal{N}}[(D_{\mathcal{X}}(G_{\mathcal{X}}(c_y, z)))^2] \\ &+ \mathbb{E}_x[(D_{\mathcal{X}}(x) - 1)^2] \\ &+ \mathbb{E}_{c_x \sim p(c_x), z \sim \mathcal{N}}[(D_{\mathcal{Y}}(G_{\mathcal{Y}}(c_x, z))^2] \\ &+ \mathbb{E}_y[(D_{\mathcal{Y}}(y) - 1)^2], \end{aligned} \tag{5}$$

where $p(c_y)$ is a marginal distribution from which c_y is sampled. Besides, we expect that the content code of a given sample should remain the same after cross-domain style transfer. This is done by minimizing the content loss ($\mathcal{L}_c$):

$$\mathcal{L}_c = \mathcal{L}_{c_x} + \mathcal{L}_{c_y} = |c_y - \hat{c}_x|_1 + |c_x - \hat{c}_y|_1, \tag{6}$$

where $|\cdot|$ is the l_1-norm, c_y (c_x) is the content code before style transfer, and $\hat{c}_x$ ($\hat{c}_y$) is the content code after style transfer. Similarly, we also expect the style code of the transferred result to be the same as the one sampled before style transfer. This is done by minimizing the style loss $\mathcal{L}_s$:

$$\mathcal{L}_s = \mathcal{L}_{s_x} + \mathcal{L}_{s_y} = |z_x - \hat{s}_x|_1 + |z_y - \hat{s}_y|_1, \tag{7}$$

where $\hat{s}_x$ and $\hat{s}_y$ are the transferred style codes, and z_x and z_y are two input style codes sampled from $\mathcal{N}(0, 1)$.

Finally, the system also incorporates self-reconstruction mechanism, as shown in the right of Fig. 1. For example, $G_{\mathcal{X}}$ should be able to reconstruct x from the latent codes (c_x, s_x) that $E_{\mathcal{X}}$ encodes. The reconstruction loss is

$$\mathcal{L}_r = \mathcal{L}_r^x + \mathcal{L}_r^y = |x - \hat{x}|_1 + |y - \hat{y}|_1, \tag{8}$$

where $\hat{x}$ and $\hat{y}$ are the reconstructed features of x and y, respectively.

RaGAN

One of our goals is to translate music pieces into the target domain with improved sound quality. To do this, we adopt the recently-proposed Relativistic average GAN (RaGAN) (Jolicoeur-Martineau 2018) as our GAN training methodology to generate high quality and stable outputs. RaGAN is different from other GAN architectures in that in the training stage, the generator not only captures the distribution of real data, but also decreases the probability that real data is real. The RaGAN discriminator is designed as

$$D(x) = \begin{cases} \sigma(Q(x) - \mathbb{E}_{x_f \sim \mathbb{Q}}\, Q(x_f)) \text{ if } x \text{ is real}, \\ \sigma(Q(x) - \mathbb{E}_{x_r \sim \mathbb{P}}\, Q(x_r)) \text{ if } x \text{ is fake}, \end{cases} \tag{9}$$

where $\sigma(\cdot)$ is the sigmoid function, Q is the layer before the sigmoid output layer of the discriminator, and x is the input data. $\mathbb{P}$ is the distribution of real data, $\mathbb{Q}$ is the distribution of fake data. x_r and x_f denote real and fake data, respectively.

Intrinsic Consistency Loss

To achieve one-to-many mapping, the MUNIT framework deprecates the cycle consistency loss that is only applicable in one-to-one settings. We needs extra ways to guarantee the robustness of the transferred features. By noticing that the multi-channel features are all derived from the mel-spectrogram with closed forms, we propose a new regularization term to guide the transferred features to be with the same closed-form relation. In other words, the intrinsic relations among the channels should remain the same after style transfer. First, the MFCC channel should remain the DCT of the mel-spectrogram:

$$\begin{aligned}\mathcal{L}_{\text{MFCC}} &= \mathcal{L}_{\text{MFCC}_u} + \mathcal{L}_{\text{MFCC}_v} \\ &= |u_{\text{MFCC}} - \text{DCT}(u_{\text{ms}})|_1 \\ &+ |v_{\text{MFCC}} - \text{DCT}(v_{\text{ms}})|_1. \end{aligned} \tag{10}$$

where u_{MFCC} is the transferred MFCC and u_{ms} is the transferred mel-spectrogram. Similar loss functions can also be

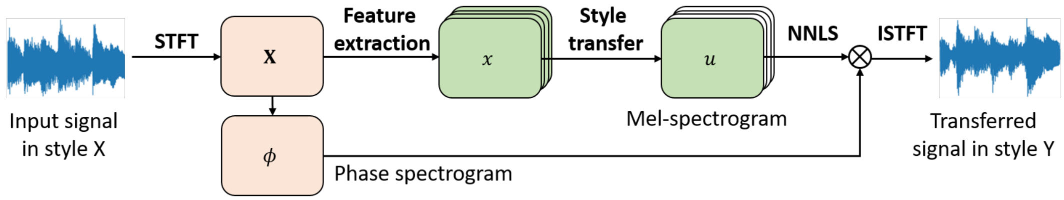

Figure 2: Illustration of pre-processing and post processing on audio signals. The power-scale spectrogram and the phase spectrogram Φ are derived from the short-time Fourier transform $\mathbf{X}$. To reconstruct the generated mel-spectrogram u_{ms}, the NNLS optimization and the original phase spectrogram Φ are used to get a stable reconstructed signal via the ISTFT.

designed for spectral difference and spectral envelope:

$$\begin{aligned}\mathcal{L}_{\Delta} &= \mathcal{L}^{u}_{\Delta} + \mathcal{L}^{v}_{\Delta} \\ &= |u_{\Delta} - \Delta u_{\mathrm{ms}}|_1 + |v_{\Delta} - \Delta v_{\mathrm{ms}}|_1 . \end{aligned} \tag{11}$$

$$\begin{aligned}\mathcal{L}_{\mathrm{env}} &= \mathcal{L}^{u}_{\mathrm{env}} + \mathcal{L}^{v}_{\mathrm{env}} \\ &= |u_{\mathrm{env}} - \mathrm{IDCT}(\mathrm{DCT}(u_{\mathrm{ms}})_{:\eta})|_1 \\ &\quad + |v_{\mathrm{env}} - \mathrm{IDCT}(\mathrm{DCT}(v_{\mathrm{ms}})_{:\eta})|_1 . \end{aligned} \tag{12}$$

That means, the transferred spectral difference (e.g., u_{Δ}) should remain as the spectral difference of the transferred mel-spectrogram (e.g., Δu_{ms}). The case of spectral envelope is also similar. The total intrinsic consistency loss is

$$\mathcal{L}_{ic} = \lambda_{\mathrm{MFCC}}\mathcal{L}_{\mathrm{MFCC}} + \lambda_{\Delta}\mathcal{L}_{\Delta} + \lambda_{\mathrm{env}}\mathcal{L}_{\mathrm{env}} \,, \tag{13}$$

and the full objective function $\mathcal{L}$ of our model is

$$\begin{aligned}&\min_{E_{\mathcal{X}},E_{\mathcal{Y}},G_{\mathcal{X}},G_{\mathcal{Y}}} \max_{D_{\mathcal{X}},D_{\mathcal{Y}}} \mathcal{L}(E_x, E_y, G_x, G_y, D_x, D_y) \\ &\qquad = \mathcal{L}_{adv} + \lambda_c\mathcal{L}_c + \lambda_s\mathcal{L}_s + \lambda_r\mathcal{L}_r + \mathcal{L}_{ic}, \end{aligned} \tag{14}$$

where λ_{adv}, λ_S and λ_{recon} are hyper-parameters to reconstruction loss.

Signal Reconstruction

The style-transferred music signal is reconstructed from the mel-spectrogram and the phase spectrogram Φ of the input signal. This is done in the following steps. First, since the mel-spectrogram $\bar{\mathbf{X}}$ is nonnegative, we can convert it back to a linear-frequency spectrogram through the mel-filterbank $\mathbf{M}$ using the nonnegative least square (NNLS) optimization:

$$\mathbf{X}^* = \arg\min_{\mathbf{X}} \|\bar{\mathbf{X}} - \mathbf{M}\mathbf{X}\|_2^2 \quad \text{subject to } \mathbf{X} \succeq 0 \,. \tag{15}$$

The resulting magnitude spectrum is therefore $\hat{\mathbf{X}} := \mathbf{X}^{*(1/\gamma)}$. Then, the complex-valued time-frequency representation $\hat{\mathbf{X}}e^{j\Phi}$ is processed by the inverse short-time Fourier transform (ISTFT), and the final audio is obtained. The process dealing with waveforms is illustrated in Fig. 2.

Implementation details

The adopted networks are mostly based on the MUNIT implementation except for the RaGAN in adversarial training. The model is optimized by adam, with the batch size being one, and with the learning rate and weight decay rate being both 0.0001. The regularization parameters in (13) and (14) are: $\lambda_r = 10$, $\lambda_s = \lambda_c = 1$, and $\lambda_{\mathrm{MFCC}} = \lambda_{\Delta} = \lambda_{\mathrm{env}} = 1$. The sampling rate of music signals is $f_s = 22.05$ kHz. The window size and hop size for STFT are 2048 and 256 samples, respectively. The dimension of the style code is 8.

Experiment and Results

In the experiments, we consider two music style transfer tasks using the following experimental data:

1. Bilateral style transfer between classical piano solo (Nocturne Complete Works performed by Vladimir Ashkenazy) and classical string quartet (Bruch's Complete String Quartet).
2. Bilateral style transfer between popular piano solo and popular guitar solo (data of both domains consists in 34 piano solos (8,200 seconds) and 56 guitar solos (7,800 seconds) covered by the pianists and guitarists on YouTube. Please see supplementary materials for details).

In brief, there are four subtasks in total: piano to guitar (P2G), guitar to piano (G2P), piano to string quartet (P2S), and string quartet to piano (S2P).

For each subtask, we evaluate the proposed system in two stages, the first being the comparison to baseline models and the second the comparison to baseline features. For the two baseline models, we consider CycleGAN (Zhu et al. 2017a) and UNIT (Liu, Breuel, and Kautz 2017), which are both competitive unsupervised style transfer networks. Note that the two baseline models allow only one-to-one mapping. For the features, we consider using mel-spectrogram only (MS), mel-spectrogram and MFCC (MC), and all four features (ALL). For simplicity, we do not exhaust all possible combinations of these settings. Instead, we consider the following five cases: CycleGAN-MS, UNIT-MS, MUNIT-MS, MUNIT-MC, and MUNIT-ALL. These cases suffice the comparison on both feature and model.

Subjective tests were conducted to evaluate the style transfer system from human's perspective. For each subtask, one input music clip is transferred using the above five settings. CycleGAN and UNIT both generate one output sample, and for MUNIT-based methods, we randomly select three style codes in the target domain and obtain three output samples. This results in a huge amount of listening samples, so we split the test into six different questionnaires, three of them comparing models and the other three three comparing features. By doing so, only one out of the three MUNIT-based output needs to be selected in a questionnaire. A participant only needs to complete one randomly selected questionnaire to finish one subjective test.

In each round, a subject first listens to the original music clip, then its three style-transferred versions using different

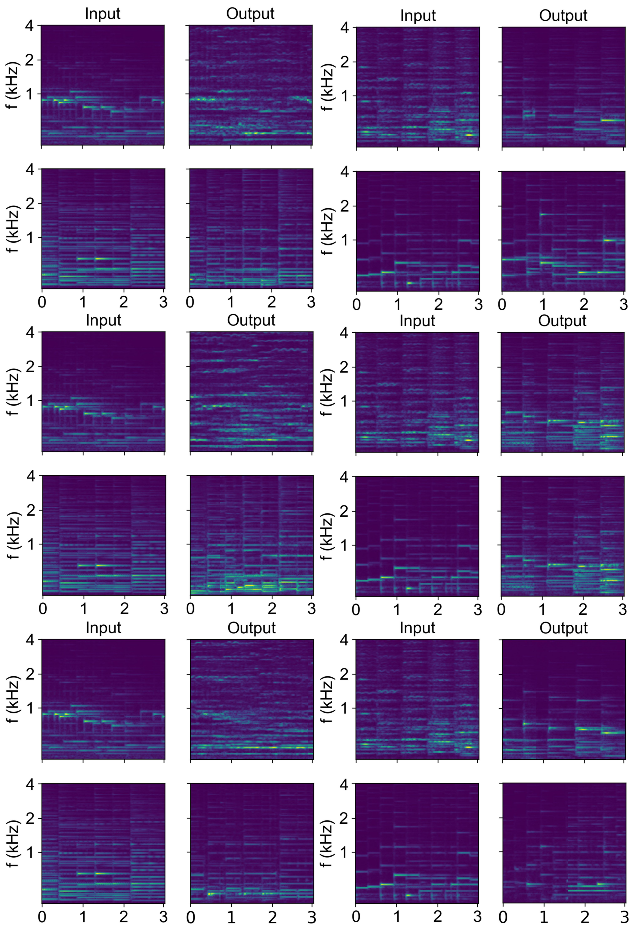

Figure 3: Comparison of the input (original) and output (transferred) mel-spectrograms for CycleGAN-MS (the upper two rows), UNIT-MS (the middle two rows), and MUNIT-MS (the lower two rows). The four subtasks demonstrated in every two rows are: P2S (upper left), S2P (upper right), P2G (lower left), and G2P (lower right).

models (i.e., CycleGAN, UNIT, MUNIT) or different features (i.e., MS, MC, ALL). For each transferred version, the subject is asked to score three problems from 1 (low) to 5 (high). The three problems are:

1. Success in style transfer (ST): how well does the style of the transferred version match the target domain,
2. Content preservation (CP): how well does the content of the transferred version match the original version, and
3. Sound quality (SQ): how good is the sound.

After the scoring process, the subject is asked to choose the best and the worst version according to her/his personal view on style transfer. This part is a preference test.

Subjective Evaluation

Table 1 shows the Mean Opinion Scores (MOS) of the listening test collected from 182 responses. First, by comparing the three models, we can see that CycleGAN performs best

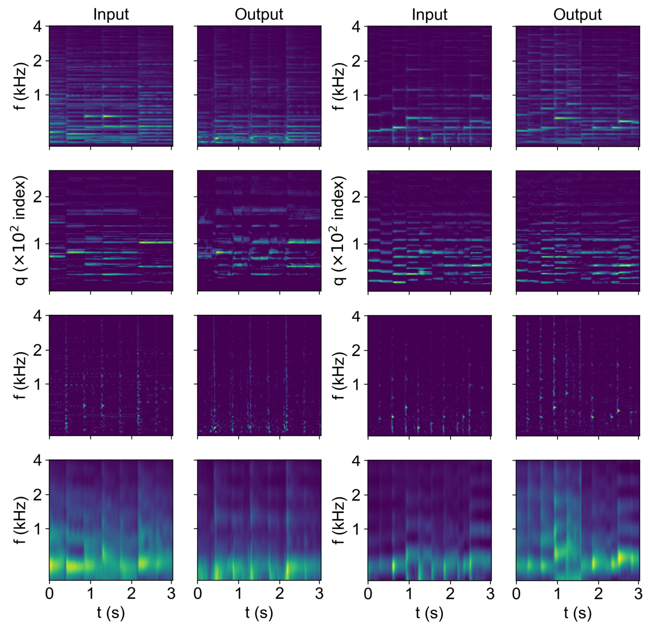

Figure 4: Illustration of the input (original) and output (transferred) feature using MUNIT-ALL of on P2G (the left two columns) and G2P (the right two columns). From top to bottom: mel-spectrogram, MFCC, spectral difference, and spectral envelope.

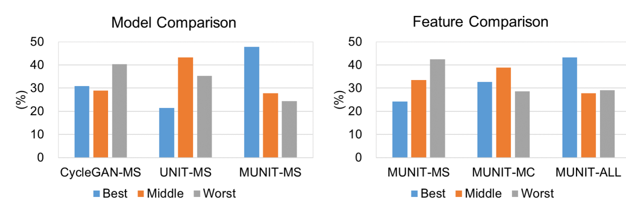

Figure 5: Results of the preference test. Left: comparison of models. Right: comparison of features. The y-axis is the ratio that each setting earns the best, middle, or the worst ranking from the listeners.

in content preservation after domain transfer, possibly because of the strength of the cycle consistency loss in matching the target domain directly at the feature level.

On the other hand, MUNIT outperforms the other two models in terms of style transfer and sound quality. Second, by comparing the features, we can see that using ALL features outperforms others by 0.1 in the average sound quality score. For content preservation and style transfer, however, the number of feature is rather insensitive. While MUNIT-based methods get the highest scores in style transfer, which shows learning a multi-modal conditional distribution better generates realistic style-transfered output, we can't see the relation between multi-channel features and style transfer quality. However, the sound quality evaluation shows that MUNIT-ALL conducts the best sound quality.

The above results indicate an unsurprising trade-off be-

Table 1: The mean opinion score (MOS) of various style transfer tasks and settings. From top to bottom: CycleGAN-MS, UNIT-MS, MUNIT-MS, MUNIT-MC, MUNIT-ALL. See the supplementary material for details about the details of evaluation.

Task		P2G			G2P			P2S			S2P			Average		
Model	Feature	ST	CP	SQ	ST	CP	SQ	ST	CP	SQ	ST	CP	SQ	ST	CP	SQ
CycleGAN	MS	2.89	**4.27**	2.56	2.66	**4.17**	2.57	2.85	3.51	2.33	3.21	**4.01**	3.10	2.90	**3.99**	2.64
UNIT	MS	2.85	4.07	2.80	2.57	3.83	2.20	2.83	**3.62**	2.28	3.39	3.90	2.88	2.91	3.85	2.54
MUNIT	MS	2.97	3.98	2.64	**3.06**	3.91	2.48	**2.88**	3.45	**2.43**	3.55	3.56	2.88	**3.12**	3.72	2.61
MUNIT	MC	3.30	4.07	**3.14**	2.80	3.56	2.42	2.77	3.32	2.27	3.47	3.44	2.92	3.09	3.60	2.69
MUNIT	ALL	**3.55**	4.12	3.13	2.95	4.02	**2.97**	2.12	3.11	1.93	**3.76**	3.70	**3.25**	3.09	3.74	**2.82**

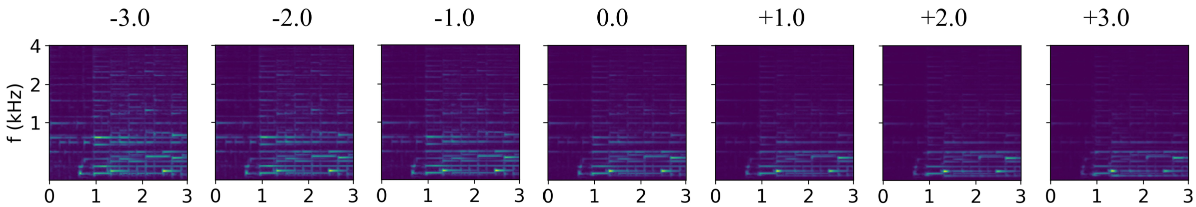

Figure 6: Converted mel-spectrograms from a piano music clip in the P2G task with the 6th dimension of the sampled style code varying from -3 to 3. The horizontal axis refers to time. Audio samples are available in the supplementary material.

tween style transfer and content preservation. The overall evaluation of listeners' preference on those music style transfer systems could be better seen from the preference test result. The results are shown in Fig. 5. For the comparison of models, up to 48% of listeners view MUNIT-MS as the best, and only 24% of listeners views it as the worst. On the other side, CycleGAN-MS gets the most "worst" votes and MNUIT-MS gets the least. For the comparison of features, 43% of the listeners view MUNIT-ALL as the best, and at the same time 42% of the listeners view MUNIT-MS as the worst. These results demonstrate the superiority of the proposed method over other baselines.

Illustration of Examples

Fig. 3 compares the input and output mel-spectrograms among different models and tasks. From the illustrations one may observe that all the models generate some characteristics related to the target domain. For example, we observe that in the P2S task, there are vibrato notes in the output, and in the P2G task, the high-frequency components are suppressed. More detailed feature characteristics can be seen in Fig. 4 where all the four features in an P2G task are shown. For the output in guitar solo style, one may further observe longer note attacks shown in the spectral difference, and less high-frequency parts in spectral envelope, both of which are indeed characteristics of guitar.

Style Code Interpolation

We then investigate how a specific dimension of the style code can affect the generation result. Fig. 6 shows a series of P2G examples with interpolated style codes. For a selected style code $z \in \mathcal{N}(0, 1)$, we linearly interpolate the 6th dimension of z, $z[6]$, with a value from -3 to 3, and generate a series of music pieces based on these modified style code. Interestingly, results show that when $z[6]$ increases, the high-frequency parts decreases. In this case, $z[6]$ can be related to some timbre features such as *spectral centroid* or *brightness*. This phenomena indicates that some of the style code elements do disentangle the characteristics of timbre.

Conclusion

We have presented a novel method to transfer a music pieces into multiple pieces in another style. We have shown that the multi-channel features in the timbre space and the regularization of the intrinsic consistency loss among them improve the sound quality of the transferred music pieces. The multimodal framework also match the target domain distribution better than previous approaches. In comparison to other style transfer methods, our proposed method is one-to-many, stable, and without the need of paired data and pre-trained model. The learned representation of style is also adjustable. These findings suggest further studies on disentangling timbre characteristics, utilizing the findings from psychoacoustics on the perceptual dimension of music styles, and the speeding up of the music style transfer system. Codes and listening examples of this work are announced online at: https://github.com/ChienYuLu/Play-As-You-Like-Timbre-Enhanced-Multi-modal-Music-Style-Transfer

References

Alluri, V., and Toiviainen, P. 2010. Exploring perceptual and acoustical correlates of polyphonic timbre. *Music Perception: An Interdisciplinary Journal* 27(3):223–242.

Aucouturier, J.-J., and Bigand, E. 2013. Seven problems that keep mir from attracting the interest of cognition and neuroscience. *Journal of Intelligent Information Systems* 41(3):483–497.

Bohan, O. B. 2017. Singing style transfer. http://madebyoll.in/posts/singing_style_transfer/.

Caclin, A.; McAdams, S.; Smith, B. K.; and Winsberg, S. 2005. Acoustic correlates of timbre space dimensions: A confirmatory study using synthetic tones. *The Journal of the Acoustical Society of America* 118(1):471–482.

Caetano, M. F., and Rodet, X. 2011. Sound morphing by feature interpolation. In *Proc. IEEE ICASSP*, 22–27.

Chen, Y.-S.; Wang, Y.-C.; Kao, M.-H.; and Chuang, Y.-Y. 2018. Deep photo enhancer: Unpaired learning for image enhancement from photographs with gans. In *CVPR*, 6306–6314.

Dai, S., and Xia, G. 2018. Music style transfer issues: A position paper. In *the 6th International Workshop on Musical Metacreation (MUME)*.

Donahue, C.; McAuley, J.; and Puckette, M. 2018. Synthesizing audio with generative adversarial networks. *arXiv preprint arXiv:1802.04208*.

Driedger, J.; Prätzlich, T.; and Müller, M. 2015. Let it bee-towards nmf-inspired audio mosaicing. In *ISMIR*, 350–356.

Gatys, L. A.; Ecker, A. S.; and Bethge, M. 2016. Image style transfer using convolutional neural networks. In *IEEE CVPR*, 2414–2423.

Goodfellow, I. J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A. C.; and Bengio, Y. 2014. Generative adversarial nets. In *NIPS*, 2672–2680.

Grey, J. M. 1977. Multidimensional perceptual scaling of musical timbres. *the Journal of the Acoustical Society of America* 61(5):1270–1277.

Gwak, J.; Choy, C. B.; Garg, A.; Chandraker, M.; and Savarese, S. 2017. Weakly supervised generative adversarial networks for 3d reconstruction. *CoRR* abs/1705.10904.

Haque, A.; Guo, M.; and Verma, P. 2018. Conditional end-to-end audio transforms. *arXiv preprint arXiv:1804.00047*.

Hosseini-Asl, E.; Zhou, Y.; Xiong, C.; and Socher, R. 2018. A multi-discriminator cyclegan for unsupervised non-parallel speech domain adaptation. *arXiv preprint arXiv:1804.00522*.

Huang, X.; Liu, M.-Y.; Belongie, S.; and Kautz, J. 2018. Multimodal unsupervised image-to-image translation. In *ECCV*.

Jolicoeur-Martineau, A. 2018. The relativistic discriminator: a key element missing from standard GAN. *CoRR* abs/1807.00734.

Kobayashi, K.; Toda, T.; Neubig, G.; Sakti, S.; and Nakamura, S. 2014. Statistical singing voice conversion with direct waveform modification based on the spectrum differential. In *INTERSPEECH*.

Larsson, G.; Maire, M.; and Shakhnarovich, G. 2016. Learning representations for automatic colorization. In *Proc. ECCV, Part IV*, 577–593.

Lartillot, O.; Toiviainen, P.; and Eerola, T. 2008. A matlab toolbox for music information retrieval. In *Data analysis, machine learning and applications*. Springer. 261–268.

Li, Y.; Liu, S.; Yang, J.; and Yang, M. 2017. Generative face completion. In *CVPR*, 5892–5900.

Liu, M.; Breuel, T.; and Kautz, J. 2017. Unsupervised image-to-image translation networks. *CoRR* abs/1703.00848.

Mao, X.; Li, Q.; Xie, H.; Lau, R. Y. K.; Wang, Z.; and Smolley, S. P. 2017. Least squares generative adversarial networks. In *ICCV*, 2813–2821.

Mor, N.; Wolf, L.; Polyak, A.; and Taigman, Y. 2018. A universal music translation network. *arXiv preprint arXiv:1805.07848*.

Peeters, G.; Giordano, B. L.; Susini, P.; Misdariis, N.; and McAdams, S. 2011. The timbre toolbox: Extracting audio descriptors from musical signals. *The Journal of the Acoustical Society of America* 130(5):2902–2916.

Siedenburg, K.; Fujinaga, I.; and McAdams, S. 2016. A comparison of approaches to timbre descriptors in music information retrieval and music psychology. *Journal of New Music Research* 45(1):27–41.

Stevens, S. S. 1957. On the psychophysical law. *Psychological review* 64(3):153.

Su, S.-Y.; Chiu, C.-K.; Su, L.; and Yang, Y.-H. 2017. Automatic conversion of pop music into chiptunes for 8-bit pixel art. In *Proc. IEEE ICASSP*, 411–415. IEEE.

Ulyanov, D., and Lebedev, V. 2016. Singing style transfer. https://dmitryulyanov.github.io/ audio-texture-synthesis-and-style-transfer/.

Välimäki, V.; González, S.; Kimmelma, O.; and Parviainen, J. 2008. Digital audio antiquing-signal processing methods for imitating the sound quality of historical recordings. *Journal of the Audio Engineering Society* 56(3):115–139.

Van Den Oord, A.; Dieleman, S.; Zen, H.; Simonyan, K.; Vinyals, O.; Graves, A.; Kalchbrenner, N.; Senior, A. W.; and Kavukcuoglu, K. 2016. Wavenet: A generative model for raw audio. In *SSW*, 125.

Verma, P., and Smith, J. O. 2018. Neural style transfer for audio spectograms. *CoRR* abs/1801.01589.

Wu, C.-W.; Liu, J.-Y.; Yang, Y.-H.; and Jang, J.-S. R. 2018. Singing style transfer using cycle-consistent boundary equilibrium generative adversarial networks. *arXiv preprint arXiv:1807.02254*.

Yu, L.; Zhang, W.; Wang, J.; and Yu, Y. 2017. Seqgan: Sequence generative adversarial nets with policy gradient. In *AAAI*, 2852–2858.

Zhang, R.; Isola, P.; and Efros, A. A. 2016. Colorful image colorization. In *Proc. ECCV, Part III*.

Zhu, J.; Park, T.; Isola, P.; and Efros, A. A. 2017a. Unpaired image-to-image translation using cycle-consistent adversarial networks. *CoRR* abs/1703.10593.

Zhu, J.; Zhang, R.; Pathak, D.; Darrell, T.; Efros, A. A.; Wang, O.; and Shechtman, E. 2017b. Toward multimodal image-to-image translation. In *NIPS*, 465–476.