Timbre-enhanced Multi-modal Music Style Transfer with Domain Balance Loss

Tsai-Jyun Fan
Department of Computer Science
National Tsing-Hua University
Hsinchu, Taiwan
Lesliefan0531@gmail.com

Chien-Yu Lu
Department of Computer Science
National Tsing-Hua University
Hsinchu, Taiwan
j19550713@gmail.com

Wei-Chen Chiu
Department of Computer Science
National Chiao Tung University
Hsinchu, Taiwan
walon@cs.nctu.edu.tw

Li Su
Institute of Information Science
Academia Sinica
Taipei, Taiwan
lisu@iis.sinica.edu.tw

Che-Rung Lee
Department of Computer Science
National Tsing-Hua University
Hsinchu, Taiwan
cherung@cs.nthu.edu.tw

***Abstract*—Style transfer of the polyphonic music recordings has always been a challenging task due to the difficulty of learning representations for both domain invariant (i.e. content) and domain-variant (i.e. style) features of the music. Although there exists prior works which employ the Multi-modal Unsupervised Image-to-Image Translation (MUNIT) framework to perform the music style transfer in an unsupervised manner and successfully provide the promising results, the gap between the transferred music recordings and the real ones is still noticeable. In order to reduce such gap, we propose and experiment several techniques for improving the transferred results, including the domain balanced loss, up-sampling, content discriminator, recycle loss, and the data scaling. We conduct extensive experiments on the tasks of bilateral style transfer among four different genres, namely: piano solo, guitar solo, string quartet, and chiptune. In evaluation, an objective testing scheme is proposed to investigate the pros and cons of all our proposed techniques, while we also design a subjective testing method for making comparison among different approaches and show that our proposed method is able to provide superior performance with respect to the prior works.**

I. Introduction

Applying style transfer on music, which aims to transfer a music recording into another realistic one as being played with different instruments, is an interesting but challenging task with great potential in practical use. In general, existing approaches on such task typically have an important assumption that a music recording can be decomposed into the content and the style attributes, hence the style transfer is attempting to modify the style of the music recording while preserving its content. However, distinguishing content and style is a challenging task due to the highly dynamic boundary between them. Traditional music style transfer methods mostly need to be performed in a supervised manner. Due to the lack of the labeled data, it becomes a restriction.

Recent approaches using deep learning methods such as the generative adversarial networks (GAN) [1] 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 and achieve higher-level mapping across domains. However, finding a proper input representation of music data for the model is not a trivial task because no clear definition are given for the music style.

This work represents an extension of [2], a recently proposed model that allows style transfer of arbitrary music recordings. The method includes timbre-enhanced data representation for the model's input, which is a combination of four timbre features. The model is based on the Multimodal Unsupervised Image-to-Image Translation (MUNIT) [3] framework to transfer the style of input music piece to another domain. Experiments show the system can achieve better performance in terms of content preservation and sound quality, and is able to find an optimal pseudo pair from non-parallel data from scratch. However, it is still hard to generate comparable timbre to music played with real world instruments.

To reduce this gap, we analyzed the problems and proposed two techniques to improve the transferred results. First, we found the previous model works well for within-domain reconstruction, but performs poorly for cross domain translation. So we proposed a new loss function, called *domain balance loss*, which constrains the discriminator from two fake data, one is from the within-domain reconstruction, and the other is from the cross-domain translation. The discriminators is forced to balance the loss from those two input. Second, the previous model cannot capture the subtle changes along the temporal axis for timbre information, including the harmonic series and the ADSR (attack, decay, sustain and release cycle). To remedy this, the model is added recycle loss, up-sampling, data scaling, and content discriminator to better capture the music features.

We conducted a series of experiments to evaluate the proposed methods. The subjective evaluation showed that with this new proposed loss, the generated results outperform [2] in terms of temporal smoothness, transferred style and sound

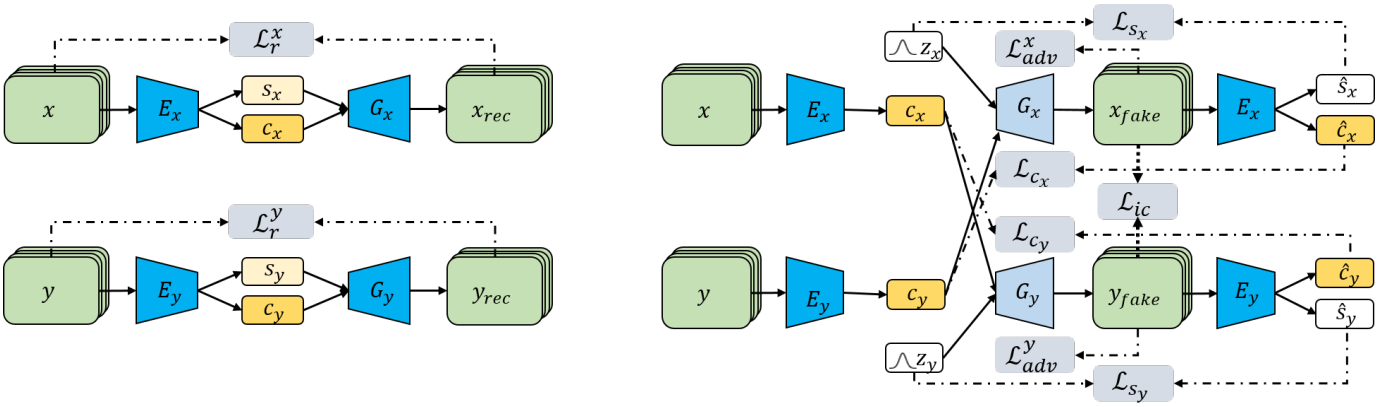

Fig. 1. The architecture of our music style transfer system. Left: within-domain reconstruction. Right: cross-domain translation.

quality. Furthermore, we used an auxiliary classifier to perform the objective evaluation of these experiments and do the further analysis. The results also indicate that the proposed methods can improve the quality of generated music.

The rest of this paper is organized as follows. Section II reviews the overall framework of [2]. Based on [2], Section III presents the proposed methods. Section IV shows the experiments and their results. Conclusion and discussion are given in the last section.

II. Timbre-enhanced style transfer: a review

Let X and Y be two domains from which data are transferred. MUNIT creates latent spaces for X and Y, each of which can be further decomposed into two subspaces: a content space and a style space. Those two domains share the same content space whereas each domain has their own style space individually. The style transfer is performed by combining the content and the style from different domains.

Our model uses four timbre features as the model inputs: 1) mel-spectrogram, 2) mel-frequency cepstral coefficients (MFCC), 3) spectral difference, and 4) spectral envelope. Those four timbre features are used as the channels of the model's input to capture the timbre information.

Figure 1 illustrates the model's architecture. Both domain X and Y have its own encoder, namely E_x, E_y and its own generator (decoder), namely G_x, G_y. The encoders convert music clips into content codes c_x, c_y and style codes s_x, s_y; the generators then take the content and style codes as input and generate another music clip that is in the style of their corresponding domain. The system has two learning paths, within-domain reconstruction and cross-domain translation. The process and the objective function for each path will be described below.

Within-domain reconstruction. To train the encoders and decoders into the inverse of each other, the system uses an self-reconstruction mechanism (music piece $\rightarrow$ latent code $\rightarrow$ music piece), as shown in the left of Fig. 1. The objective function calculates the reconstruction loss between input music pieces and reconstructed music pieces.

$$\mathcal{L}_r = \mathcal{L}_r^x + \mathcal{L}_r^y = |x - x_{rec}|_1 + |y - y_{rec}|_1 \tag{1}$$

where $|\cdot|$ is the $l1$-norm, x_{rec}, y_{rec} are the reconstructed features of x and y, respectively. The reconstruction path of x_{rec}, y_{rec} can be written as:

$$x_{rec} = G_x(E_x^c(x), E_x^s(x)) \tag{2}$$
$$y_{rec} = G_y(E_y^c(y), E_y^s(y)) \tag{3}$$

Cross-domain translation. Different from the within-domain reconstruction that the style codes are encoded from the input data, the generators take the style codes $z_x, z_y \in N(0,1)$, randomly sampled from a Gaussian distribution and the content codes, encoded from another domain's music piece. This means that they preserve the content of the other domain's input and translate its style to their own domain. The across domain translation is achieved through the generative adversarial approach. To fool the discriminators D, the encoders and generators need to cooperate and match the distribution of the transferred music pieces to the target style distribution, while the discriminators try to distinguish between the transferred results and the real data. That means, if the model can generate results which cannot be distinguished by the discriminators from the ones in the real target domain, then it has successfully captured the distribution of the target style.

The relativistic average LSGAN (RaLSGAN) [4] is used as the GAN loss to generate better results in sound quality. Different from other GAN loss in the training stage, the generator does not only capture the distribution of real data, but also decreases the probability that real data is real. The loss function for the generator can be represented as:

$$\begin{aligned}\mathcal{L}_{adv}^G &= \mathcal{L}_{adv}^{G_x} + \mathcal{L}_{adv}^{G_y} \\ &= \mathbb{E}_x[((D_X(x) - \mathbb{E}_{x_{fake}} D_X(x_{fake})) + 1)^2] \\ &+ \mathbb{E}_{x_{fake}}[((D_X(x_{fake}) - \mathbb{E}_x D_X(x)) - 1)^2] \\ &+ \mathbb{E}_y[((D_Y(y) - \mathbb{E}_{y_{fake}} D_Y(y_{fake})) + 1)^2] \\ &+ \mathbb{E}_{y_{fake}}[((D_Y(y_{fake}) - \mathbb{E}_y D_Y(y)) - 1)^2]\end{aligned} \tag{4}$$

and for the discriminator:

$$\begin{aligned}\mathcal{L}_{adv}^{D} &= \mathcal{L}_{adv}^{D_x} + \mathcal{L}_{adv}^{D_y} \\ &= \mathbb{E}_x[((D_X(x) - \mathbb{E}_{x_{fake}} D_X(x_{fake})) - 1)^2] \\ &+ \mathbb{E}_{x_{fake}}[((D_X(x_{fake}) - \mathbb{E}_x D_X(x)) + 1)^2] \\ &+ \mathbb{E}_y[((D_Y(y) - \mathbb{E}_{y_{fake}} D_Y(y_{fake})) - 1)^2] \\ &+ \mathbb{E}_{y_{fake}}[((D_Y(y_{fake}) - \mathbb{E}_y D_Y(y)) + 1)^2]\end{aligned} \tag{5}$$

where x_{fake}, y_{fake} are the transferred music pieces that are in the distribution of fake data in X and Y domains, respectively. The translation path from domain X to domain Y ($x \rightarrow y_{fake}$) and from domain Y to domain X ($y \rightarrow x_{fake}$) can be written as:

$$y_{fake} = G_y(E_x^c(x), z_y) \text{ and } x_{fake} = G_x(E_y^c(y), z_x) \tag{6}$$

Additionally, the model uses the reconstruction loss in the latent space (latent code $\rightarrow$ music piece $\rightarrow$ latent code) which ensures the information of content and style are still preserved after the cross-domain translation. Below illustrate the content($\mathcal{L}_c$) and the style reconstruction loss ($\mathcal{L}_s$) respectively:

$$\mathcal{L}_c = \mathcal{L}_{c_x} + \mathcal{L}_{c_y} = |c_x - \hat{c}_y|_1 + |c_y - \hat{c}_x|_1 \tag{7}$$

where c_x (c_y) is the content code before style transfer and $\hat{c}_x$ ($\hat{c}_y$) is the content code encoded back from the transferred results.

$$\mathcal{L}_s = \mathcal{L}_{s_x} + \mathcal{L}_{s_y} = |z_x - \hat{s}_x|_1 + |z_y - \hat{s}_y|_1 \tag{8}$$

where z_x (z_y) is the style code random sampled from $N(0, 1)$ in a Gaussian distribution and $\hat{s}_x$ ($\hat{s}_y$), is the style code encoded back from the transferred results. The full objective function $\mathcal{L}$ of our model is

$$\mathcal{L} = \lambda_{adv}\mathcal{L}_{adv} + \lambda_c\mathcal{L}_c + \lambda_s\mathcal{L}_s + \lambda_r\mathcal{L}_r + \lambda_{ic}\mathcal{L}_{ic} \tag{9}$$

where $\mathcal{L}_{adv}$ is the combination of $\mathcal{L}_{adv}^G, \mathcal{L}_{adv}^D$, and $\mathcal{L}_{ic}$ is the intrinsic consistency loss for keeping the consistency among the channel-wise features in the target domain thus improve the sound quality, we don't describe the detail of this loss since it's less relevant to our improvement in this work. Readers are encouraged to refer to more details of the method in [2].

III. Proposed Methods

A. Domain Balance Loss

Several problems for the timbre-enhanced style transfer remain. First, the generated timbre is different from that generated by real instruments and the sound quality also can be improved. Second, when combining the transferred music clips with the raw wav file, adjacent music clips along the timeline often loss the continuity and consistency. In other words, two music clips adjacent to each other in a song may sound discontinuous. For example, they might sound like the same instruments but played with different brands. Notice that during the inference time, the style code is fixed for the whole song so all the music clips that compose the song should be in a consistent style.

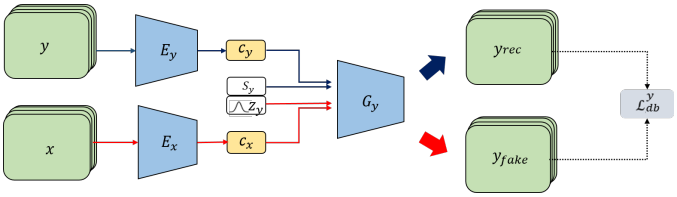

Fig. 2. The view of the two learning paths of our generation model, we omit G_x ($\mathcal{L}_{db}^x$) here for simplicity. The figure can be spilt into two parts: up (blue arrow): within-domain reconstruction, down (red arrow): cross-domain translation.

The problems described above are not noticeable when the results are generated within the same domain, or even with a randomly sampled style code. It means that the model actually is able to perform well when the generation doesn't involve the cross-domain translation. Figure 2 gives an illustration of the difference between two generation paths (within-domain, cross-domain) of our model from G_y perspective. The generator G_y needs to take a content code c_y/c_x and a style code s_y/z_y as its input. Since replacing s_y with z_y actually doesn't cause too much loss in the quality of within-domain reconstructed results, we can infer that the reason for the difference in quality between the results of two paths is that whether the encoder $E_y(E_x)$ that encodes the input into $c_y(c_x)$ belongs to the same domain as the decoder. For example, when reconstructing the input within domain, G_y decodes the content code c_y that encodes by E_y (same domain). On the contrary, when transferring the input across domains, it decodes the content code c_x that encodes by E_x (different domain).

We expect that the generator should decode to results with similar quality no matter the content code it takes is encoded from which domain. We assume this unbalance situation is because that we trained the within-domain reconstruction with $\mathcal{L}_r$, which is a strong constrain loss with ground truth and easier to learn than the adversarial loss that is used to train the cross-domain path, and thus accidentally dominate the main learning direction of our model.

To fix this problem, we added a constrain loss to balance the learning of two paths, named *domain balance loss* ($\mathcal{L}_{db}$). Figure 2 specified the main idea of our approach. According to the property of the GAN loss (RaLSGAN), the discriminators take two inputs, one real data and one fake data, and output the probability that the given real data is more realistic than the fake data. Now, the new constrain forces discriminators take two fake data: one from within-domain reconstruction; the other from cross-domain translation. With the constrain, the discriminators will give them similar probability of being fake or real data, since they are both belong to the distribution of fake data and none of them should be far more real than the other. Consequently, the reconstruction path and the translation path can map their outputs to the similar distribution, balancing the learning of the encoders, and alleviate the problem of the unbalance learning of the two paths.

The proposed loss is shown below.

$$\begin{aligned}\mathcal{L}_{db} &= \mathcal{L}_{db}^{x} + \mathcal{L}_{db}^{y} \\ &= (\mathbb{E}_{x_{rec}} D(x_{rec}) - \mathbb{E}_{x_{fake}} D(x_{fake}))^2 \\ &+ (\mathbb{E}_{y_{rec}} D(y_{rec}) - \mathbb{E}_{y_{fake}} D(y_{fake}))^2\end{aligned} \tag{10}$$

where x_{rec}, y_{rec} are generated from within-domain reconstruction path and x_{fake}, y_{fake} are generated from cross-domain translation path. The proposed new loss can be easily integrated with our model by appending it to the original objective function, so our new objective function can be written as:

$$\mathcal{L}_{new} = \mathcal{L}_{ori} + \lambda_{db}\mathcal{L}_{db} \tag{11}$$

The model architecture is modified from the MUNIT architecture. We changed the activation function of the last layer of the decoder from Tanh to ReLU since the number in the music features shouldn't contain negative numbers. The model is optimized by ADAM [5], with the batch size one, and the learning rate and weight decay rate are both 0.0001. We set the regularization parameters of the loss functions to: λ_{adv} = 1, $\lambda_c = \lambda_s$ = 1, λ_r = 10, λ_{ic} = 1 and λ_{db} = 0.5.

B. ADSR Enhancement

We have observed that the transferred outputs do not truly sound like real instruments, and one reason is because that the ADSR (attack, decay, sustain and release) curves of the transferred notes are not well captured by the model. Typically, ADSR is related to both long-term dependency (e.g., sustained strings) and short-term dependency (e.g., acute and strong attacks of percussion) in audio signals. To better capture the ADSR behaviors of the target styles, there are two possible strategies: 1) using the data representation with a higher resolution (i.e. enhancement of local information), and 2) extending the temporal range of the data representation (i.e. enhancement of global information). We therefore consider three methods to fulfill these purposes, which are up-sampling, recycle loss, and data scaling.

Up-sampling. In order to let the model has the ability to reconstruct the details of music data representations, we cropped each input into sub-segments along the temporal axis (i.e. the x-axis of data representation), and trained local discriminators for both source and target domain to discriminate the real sub-segments from the fake ones. Then, bilinear interpolation is performed to re-sample the sub-segment up to the size of the original input data along the temporal axis. Through this auxiliary adversarial setting, we expect that the generators are guided to generate results with the right details, and the local discriminators can easily discriminate its input since up-sample the cropped data with the wrong detail would increase the unreality of the data.

Recycle Loss. Recycle-GAN indicates that the use of temporal information in videos can provide more constraints to the optimization for transforming one domain to another, by applying a novel loss function named Recycle Loss [6]. Recycle Loss is a revision of the cycle loss (i.e. reconstruction loss) function used in Cycle-GAN. In comparison to cycle loss, recycle loss adopts one more predictor to predict the forthcoming video frame from the previous video frames. It then calculates the reconstruction loss with the predicted video frame. Similar to videos, our model includes the Recycle Loss by replacing the original fake images in the reconstruction part of the total objective function ($\mathcal{L}_r$, $\mathcal{L}_c$, $\mathcal{L}_s$) with the predicted fake images.

Data scaling. Another strategy to create different resolutions of data representation is to use different hop sizes for STFT (short-time Fourier transform). A smaller hop size can produce smaller temporal time grids of data representation, i.e. the data points are denser and therefore contain more precise local information. We assume that our choice of hop size may be too big and cause the loss of data information. Our original hop size is set to 256, to increase the detail information in our data, we consider smaller hop size (128, 96) to see that whether this change can get any improvement of our results.

Content discriminator. In addition to focus on how to let our model capture the ADSR of timbre, we also made an assumption that the drop of performance in cross-domain translation compared to within-domain reconstruction and is due to the inefficiency of disentanglement, which means that the content codes still contain style information in it. Inspired by recent works including DRIT [7] and a voice conversion task [8], we also consider adding a content discriminator on the latent space to distinguish the extracted content codes between the two domains, and to facilitate the ablation of domain information in content codes. The architecture of the content discriminator is based on the one proposed by DRIT [7].

IV. Experiment and Results

We consider three music style transfer tasks. The tasks and used data sets are described as below.

1) Bilateral style transfer between 8,200 seconds of popular piano solo and 7,800 seconds of popular guitar solo covered by the pianists and guitarists on YouTube.
2) Bilateral style transfer between 6,701 seconds of classical piano solo and 4,796 seconds of classical string quartet.
3) Bilateral style transfer between 8,200 seconds of popular piano solo (same as task1) and 9210 seconds of popular 8-bit (chip-tune) covered by the pianists and made by 8-bit music player on YouTube, respectively.

In other words, there are six sub-tasks in total: piano to guitar (P2G), guitar to piano (G2P), piano to string quartet (P2S), string quartet to piano (S2P), piano to 8-bit (P2B), and 8-bit to piano (B2P).

The above-mentioned tasks are evaluated in two approaches: objective test. and subjective tests. Since asking human to compare all parameter settings mentioned in Section III in the subjective test is ineffective, an object evaluation is required to verify the parameter settings. In the subjective test, we evaluated the proposed model with domain balance loss (denoted

as MUNIT-DB) by comparing it with two baseline models, MUNIT in [2] and Recycle-GAN [6], for each sub-task. Recycle-GAN is a widely-used, competitive unsupervised style transfer network, which improve the Cycle-GAN [9], which overall optimization process is focused on reconstructing the input when transferring style of videos. Recycle-GAN then successfully improved the quality of the transferred videos by considering the temporal information of data. Note that unlike MUNIT and the proposed method, Recycle-GAN allows only one-to-one mapping.

A. Objective Evaluation

In the objective test, we trained an instrument classifier that classify among three classes of instrument (i.e. guitar, piano, and strings). The main idea of the objective test is that a higher classification accuracy should be obtained for the outputs of a better style transfer model on the target style labels. This provides a simple yet practical way to evaluate the performance of style transfer models without subjective tests, and can be helpful when there are many model parameters to be fine-tuned and compared. In this work, the classifier network structure is formed by four 4×4 convolution layers, followed by a 1×1 convolution layer whose output dimension is 1 for dimension reduction. The output dimensions of the former 4 layers are 64, 128, 256, 512, respectively. Lastly, a fully connected layer is added to output three dimensions which represents the three output instrument classes.

To train our style transfer model and instrument classifier with independent training sets, a large amount of data is needed. To do this, we use MIDI files with an audio synthesizer from music21 toolkit to generate music data of the three instruments. We then split the generated data into two training sets and one testing set. One of the two training sets is used for training our style transfer model, and the other one is for training the instrument classifier. The testing data are used as the input of the style transfer models, and the transferred results are evaluated with the instrument classifier. Since the data are all synthetic, the instrument classifier can achieve 99.95 % accuracy (i.e. 0.05% error rate) on the original music data without style transfer; see Table I.

Table I lists the classification error rates (in %) of the proposed style transfer methods under various parameter settings. Here, MUNIT denotes the baseline model [2]; MUNIT-DB (this work) is the MUNIT model with the domain balance loss; MUNIT-RL is the one including recycle loss [6]; MUNIT-US is the one using up-sampling; MUNIT-H128 and MUNIT-H96 are the ones applying data scaling (i.e. one with hop size of 128 and the other with hop size of 96); MUNIT-CD is the one with content discriminators and and RCGAN is the one using Recycle-GAN. As we can see, MUNIT-DB achieves the lowest error rate among all. This is also consistent with the results of subjective tests that MUNIT-DB receives the highest score from the users; see the result of subjective evaluation. It should also be noted that MUNIT-RL and MUNIT-US achieves slight though insignificant improvement in comparison to MUNIT. MUNIT-H128, MUNIT-H96, and MUNIT-CD get worse results than MUNIT.

From the results above, the optimal setting of the style transfer models can be obtained without conducting subjective tests: we find that reducing the hop size for STFT will only make the transferred results worse. By fixing the number of frames in each input, choosing different hop sizes actually results in a trade off between enhancing local information and enhancing global information; this is the reason why MUNIT-H128 and MUNIT-H96 do not improve the results. On the other hand, MUNIT-RL and MUNIT-US, the methods which manage to enhance global and local information, respectively, both result in performance improvement compared to MUNIT, though the extent of improvement is limited. The result of MUNIT-CD shows that the adversarial way cannot truly lead the model to achieve better style transfer. as it tends to add noises in content codes to fool the content discriminator. The result of RCGAN shows relatively higher error rate than MUNIT, indicates that Recycle-GAN has weaker performance in ST metric which will be further confirmed by subjective evaluation in next part. Based on the classification results, we consider using only MUNIT-DB in the subjective test.

B. Subjective Evaluation

In each sub-task, the subjects are asked to listen to the original music clip, as well as three versions of that music clip after style transfer using three different models: MUNIT-DB, MUNIT, and Recycle-GAN.

For each transferred music clip, we asked the subjects to score the three metrics from 1 (low) to 7 (high) so that we can get the mean opinion score (MOS) and do the further evaluation. The three metrics are:

1) Temporal smoothness (TS): the continuity and consistency along the temporal axis of the transferred music,
2) Success in style transfer (ST): how well does the style of the transferred music match the target domain (target instrument), and
3) Sound quality (SQ): how good does the transferred music sounds.

We also conducted the preference test by asking the subjects to choose the best and the worst version according to their personal view on style transfer. These questions are listed into an online questionnaire. The questionnaire is distributed through social media.

We received 152 responses, where 32% of them reported themselves have music related experience, or are familiar with the timbre of our experimental instruments. Table II summarizes the mean opinion scores (MOS) of the listening test in the above three metrics. The following observations are made. In all the three metrics, our work (MUNIT-DB) gets the highest score on average, but not in all the sub-tasks. As can be seen, when comparing MUNIT-DB with MUNIT, MUNIT-DB gets better score for all the sub-tasks. Comparing MUNIT-DB with Recycle-GAN, MUNIT-DB performs the best in style transfer under any circumstances, but Recycle-GAN outperforms MUNIT-DB in temporal smoothness or sound

TABLE I
THE RESULTS OF OBJECTIVE EVALUATION BASED ON CLASSIFICATION.

Task	Original	MUNIT	MUNIT-DB	MUNIT-RL	MUNIT-US	MUNIT-H128	MUINT-H96	MUNIT-CD	RCGAN
Error Rate(%)	0.05	0.46	0.11	0.37	0.24	1.66	7.36	6.97	2.23

TABLE II
THE MEAN OPINION SCORE (MOS) OF VARIOUS STYLE TRANSFER TASKS AND SETTINGS.

Task	P2G			G2P			P2S			S2P			P2B			B2P			Average		
Model	TS	ST	SQ	TS	ST	SQ	TS	ST	SQ	TS	ST	SQ	TS	ST	SQ	TS	ST	SQ	TS	ST	SQ
Recycle-GAN	5.29	3.95	3.99	5.04	3.26	**3.91**	**4.97**	3.45	**4.00**	5.07	4.52	4.28	3.67	3.95	3.03	5.36	3.86	**4.65**	4.90	3.83	3.98
MUNIT	5.36	4.30	4.25	5.13	3.05	3.83	3.69	2.87	2.53	5.00	4.86	4.03	3.93	3.96	2.95	5.16	3.99	4.08	4.71	3.84	3.61
MUNIT-DB	**5.50**	**4.72**	**4.45**	**5.19**	**3.40**	3.86	4.43	**3.55**	3.28	**5.49**	**5.24**	**4.45**	**4.18**	**4.30**	**3.35**	**5.41**	**4.55**	4.59	**5.03**	**4.29**	**4.00**

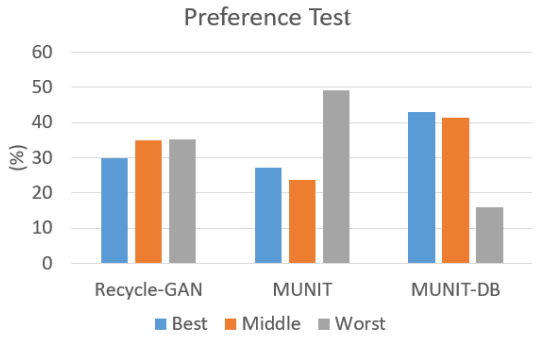

Fig. 3. Results of the preference test. The y-axis is the ratio that each setting earns the best, middle, or the worst ranking from the listeners

quality. However, Recycle-GAN sometimes gets a worse score than MUNIT especially on the style transfer metric. Above results indicate that there is a trade-off between style transfer and two other metrics in some sub-tasks when Recycle-GAN is compared with our work.

According to these observations, the Recycle-GAN does increase the temporal smoothness of sound in some transfer sub-tasks by adding temporal information to the model and may achieve better sound quality, but often with the side effect of making the sound less realistic. On the other hand, our method can always perform well in style transfer, and also brings some improvement on two other metrics compare to MUNIT. In general, our method is the most stable when considering all those three metrics and performs the best in style transfer metric.

We can further confirm this claim through the preference test, as shown in Figure 3, in which there are 42.87% of listeners think that MUNIT-DB performs the best and only 27.18%, 29.95% listeners choose MUNIT, Recycle-GAN as the best, respectively. We can also confirm our theory from the worst aspect that only 15.78% listeners choose MUNIT-DB as the worst, which is far less than MUNIT (49.12 %) and Recycle-GAN (35.1%). Also, in each sub-task, MUNIT-DB never gets the "worst" votes like MUNIT and Recycle-GAN. The analysis and results above demonstrate the stability and superiority of the proposed method over other two baselines.

V. CONCLUSIONS

In this paper we present a simple yet efficient loss (i.e. domain balance loss) which successfully improves the previous work built upon MUNIT for the efficacy of music style transfer, and shows better results in terms of the temporal smoothness and sound quality. The observations found in our extensive experiments and the thorough studies on several techniques, together with our evaluation schemes, are able to provide valuable reference for the following researches. Codes and listening examples of this work are announced online at: https://github.com/LeslieFan0531/Play-As-You-Like-Timbre-Enhanced-Multi-modal-Music-Style-Transfer. For future work, in addition to continuously improving the quality of music style transfer, we attempt to also apply the idea of our proposed domain balance loss on other tasks (e.g. image or video style transfer) for further investigating its generalizability and the practical potentials.

REFERENCES

[1] 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 (NIPS)*, 2014.

[2] C.-Y. Lu, M.-X. Xue, C.-C. Chang, C.-R. Lee, and L. Su, "Play as you like: Timbre-enhanced multi-modal music style transfer," in *AAAI Conference on Artificial Intelligence (AAAI)*, 2019.

[3] X. Huang, M.-Y. Liu, S. Belongie, and J. Kautz, "Multimodal unsupervised image-to-image translation," in *European Conference on Computer Vision (ECCV)*, 2018.

[4] A. Jolicoeur-Martineau, "The relativistic discriminator: a key element missing from standard gan," in *International Conference on Learning Representations (ICLR)*, 2019.

[5] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations (ICLR)*, 2014.

[6] A. Bansal, S. Ma, D. Ramanan, and Y. Sheikh, "Recycle-gan: Unsupervised video retargeting," in *European Conference on Computer Vision (ECCV)*, 2018.

[7] H.-Y. Lee, H.-Y. Tseng, J.-B. Huang, M. Singh, and M.-H. Yang, "Diverse image-to-image translation via disentangled representations," in *European Conference on Computer Vision (ECCV)*, 2018.

[8] J.-C. Chou, C.-C. Yeh, H.-Y. Lee, and L.-S. Lee, "Multi-target voice conversion without parallel data by adversarially learning disentangled audio representations," in *Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2018.

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