MUSIC PHRASE INPAINTING USING LONG-TERM REPRESENTATION AND CONTRASTIVE LOSS

Shiqi Wei[1,2], *Gus Xia*[2], *Yixiao Zhang*[3], *Liwei Lin*[2], *Weiguo Gao*[4,1]

[1] School of Data Science, Fudan University
[2] Music X lab, New York University Shanghai
[3] Centre for Digital Music, Queen Mary University of London
[4] School of Mathematical Sciences, Fudan University

ABSTRACT

Deep generative modeling has already become the leading technique for music automation. However, *long-term* generation remains a challenging task as most methods fall short in preserving a natural structure and the overall musicality when the generation scope exceeds several beats. In this study, we tackle the problem of long-term, phrase-level symbolic melody inpainting by equipping a sequence prediction model with phrase-level representation (as an extra condition) and contrastive loss (as an extra optimization term). The underlying ideas are twofold. First, to predict phrase-level music, we need phrase-level representations as a better context. Second, we should predict notes and their high-level representations simultaneously, while contrastive loss serves as a better target for abstract representations. Experimental results show that our method significantly outperforms the baselines. In particular, contrastive loss plays a critical role in the generation quality, and the phase-level representation further enhances the structure of long-term generation.[1]

***Index Terms*—** Music Inpainting, Contrastive Learning, Representation Learning, Deep Music Generation

1. INTRODUCTION

In recent years, deep generative models have achieved promising progress in the field of symbolic music generation [1–3]. In particular, *music inpainting* task [4–7] draws lots of research attention due to its great practical value in human-computer music co-creation [8]. The general setting is that human composers create some parts of a piece, while the algorithm inpaints (or infills) the rest. However, long-term generation remains a challenging task. When the inpainting scope exceeds several beats, current methods cannot yet preserve a natural structure and the overall musicality.

To solve the problem, we resort to music representation learning [1, 9–13]. The main idea is that well-learned music representations can greatly benefit generation tasks since music creation is not merely a series of note-by-note decisions but also involves the natural flow of high-level abstractions (representations). Several pioneer inpainting studies [6, 7] are following such direction and predict bar-level representations rather than notes. Compared to the existing studies, the contributions of this paper are:

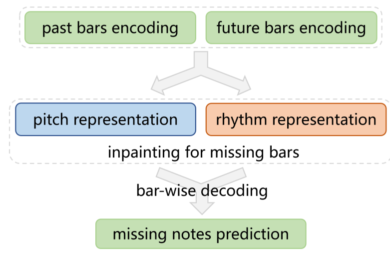

Fig. 1: A general system diagram of the proposed system.

- **Significantly longer generation**: we contribute the first *phrase-level* music inpainting model, in which the generation scope is considerably longer (16 beats, or 4 bars in $\frac{4}{4}$ meter).
- **Long-term compact contexts**: to inpaint phrase-level music, we incorporate phrase-level representations as an auxiliary context. As far as we know, this is also the first study that successfully leverages the pre-trained phrase representation for prediction purpose.
- **Dual-prediction scheme**: We predict notes and bar-level representations in parallel, during which *contrastive loss* [14, 15] is designed for latent representations while ordinary cross-entropy loss is applied to individual notes.

Fig. 1 shows the overall system diagram. In this paper, we focus on pop/folk songs that are four-phrase long and always inpaint the missing melody of the *third* phrase given other parts of the lead sheet (melody and the underlying chords). In addition, inspired by [7], we estimate pitch and rhythm representations separately and then rely on EC^2-VAE [9] to

[1]The code is released at https://github.com/SqWei17/music_phrase_inpainting.

integrate the predicted pitch and rhythm representations and reconstruct the missing notes. Experimental results show that the dual scheme with contrastive loss plays a key role in high-quality inpainting, dramatically outperforming the baseline. Moreover, phrase-level context further enhances the music structure.

2. METHODOLOGY

The model contains three parts: 1) pre-trained EC^2-VAE encoders to provide context representations in both bar and phrase levels, 2) a tailored forward-backward inpainting model, and 3) contrastive loss for representation estimation.

2.1. Pre-trained EC^2-VAE Encoders

EC^2-VAE [9] is an off-the-shelf music representation learning model. It relies on an extra rhythm decoder (with explicit rhythm loss) to extract disentangled pitch and rhythm representations for music within two bars. Its follow-up study [13] extends the model to a hierarchical EC^2-VAE and successfully extracts phrase-level pitch and rhythm representations.

We adopt the pre-trained encoders to extract both bar-level and phrase-level representations. Formally, given a 4-phrase melody $M = \{m_1, m_2, m_3, m_4\}$, the target of the model is to infill the missing phrase m_3. We use $c_i^u (i \in \{1, 2, 3, 4\})$ to denote the representation of the *i*-th phrase and use $z_{i,j}^u (i, j \in \{1, 2, 3, 4\})$ for the representation of the *j*-th bar in the *i*-th phrase, where $u \in \{\text{p, r}\}$ (p for pitch and r for rhythm).

2.2. Disentangled Representation Inpainting

As shown in Fig. 2, the proposed bi-directional inpainting model comprises a forward-backward GRU module and a estimation module. The model takes disentangled pitch and rhythm representations generated by pre-trained EC^2-VAE encoders as input and then inpaint the missing phrase m_3.

2.2.1. *Forward-backward GRU Module*

Let $z_{\text{past}}^u = \{z_{1,1}^u, \cdots, z_{1,4}^u, z_{2,1}^u, \cdots, z_{2,4}^u\}$ and $z_{\text{future}}^u = \{z_{4,1}^u, \cdots, z_{4,4}^u\}$. Conditioned on the phrase-level representation c_4^u, the forward GRU module encodes z_{past}^u into a forward contextual representation:

$$\overrightarrow{h}^u = \text{ForwardGRU}_u(z_{\text{past}}^u, c_4^u) \tag{1}$$

Similarly, conditioned on $c_{1:2}^u$, the backward GRU module for u encodes z_{future}^u into a backward contextual representation:

$$\overleftarrow{h}^u = \text{BackwardGRU}_u(z_{\text{future}}^u, c_{1:2}^u) \tag{2}$$

Here, we condition phase-level representations on GRUs by feeding them as the initial hidden vectors of GRUs.

2.2.2. *Estimation Module*

The estimation module comprises linear layers and a EC^2-VAE decoder. The forward-backward GRUs mentioned in the previous section and linear layers take $\overrightarrow{h}^u$ and $\overleftarrow{h}^u$ as input to generate bar-level disentangled representations of m_3, $\hat{z}_{3,j}^u (u \in \{p, r\}, j \in \{1, 2, 3, 4\})$. Finally, the module inpaints the missing phrase by decoding $\hat{z}_{3,j}^u$ using a EC^2-VAE decoder.

Assuming $m_3 = \{x_1, x_2, x_3, x_4\}$, we denote the predicted melody sequence as $\hat{m}_3 = \{\hat{x}_1, \hat{x}_2, \hat{x}_3, \hat{x}_4\}$. Then the forward bar-level disentangled representation $\overrightarrow{h}_{3,j}^u (u \in \{p, r\}, j \in \{1, 2, 3, 4\})$ for u at bar-j in m_3 generated by the forward GRU is:

$$\widetilde{z}_{3,j}^u = \text{EC}^2\text{-VAE-Encoder}(\hat{x}_j, \text{chord}_j) \tag{3}$$

$$\overrightarrow{v}_{3,j}^u = \begin{cases} \text{ForwardGRU}_u(\widetilde{z}_{3,j-1}^u, \overrightarrow{v}_{3,j-1}^u) & \text{, if } j > 1 \\ \overrightarrow{h}^u & \text{, if } j = 1 \end{cases} \tag{4}$$

$$\overrightarrow{h}_{3,j}^u = w_1^{u\mathsf{T}} \overrightarrow{v}_{3,j}^u \tag{5}$$

where w_1^u is a linear layer and chord_j is the chord of x_j. The backward bar-level disentangled representation $\overleftarrow{h}_{3,j}^u$ for u at bar-j is:

$$\overleftarrow{v}_{3,j}^u = \begin{cases} \text{BackwardGRU}_u(\overleftarrow{h}_{3,j+1}^u, \overleftarrow{v}_{3,j+1}^u) & \text{, if } j < 4 \\ \overleftarrow{h}^u & \text{, if } j = 4 \end{cases} \tag{6}$$

$$\overleftarrow{h}_{3,j}^u = w_2^{u\mathsf{T}} \overleftarrow{v}_{3,j}^u \tag{7}$$

where w_2^u is a linear layer. We concatenate $\overrightarrow{h}_{3,j}^u$ and $\overleftarrow{h}_{3,j}^u$ together and encodes them into the target bar-level disentangled representation $\hat{z}_{3,j}^u$ using a linear layer W and then inpaint the bar-j of the missing phrase:

$$\hat{z}_{3,j}^u = W^{\mathsf{T}}[\overrightarrow{h}_{3,j}^u \; \overleftarrow{h}_{3,j}^u], u \in \{\text{p, r}\} \tag{8}$$

$$\hat{x}_j \sim \text{EC}^2\text{-VAE-Decoder}(\hat{z}_{3,j}^{\text{p}}, \hat{z}_{3,j}^{\text{r}}, \text{chord}_j) \tag{9}$$

We use the cross-entropy loss $\mathcal{L}_{\text{CE}}$ between x_j and the distribution of $\hat{x}_j$ to update the weights of the forward-backward GRU module and estimation module.

2.3. Contrastive Loss

We apply contrastive loss $\mathcal{L}_{\text{CL}}$ on $\hat{z}_{3,j}^u$ to encourage the inpainted representations to inherit some abstract information of the original pieces. We regard the bar-level disentangled representation of ground truth pieces generated by the pre-trained 1-bar EC^2-VAE encoder as positive samples of $\hat{z}_{3,j}^u$. More specifically, we regard $z_{3,j}^u$ as positive samples and $z'^u = \{z_1'^u, .., z_K'^u\}$ as negative samples, where z'^u is bar-level disentangled representations of other songs.

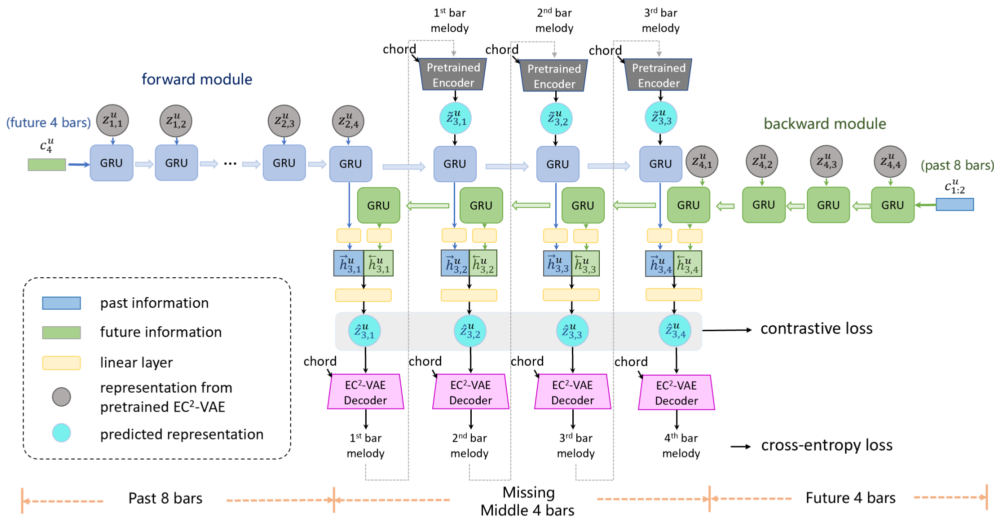

Fig. 2: The bi-directional inpainting model. u is the disentangled factor that $u \in \{\mathrm{p}, \mathrm{r}\}$.

The contrastive loss $\mathcal{L}_{\mathrm{CL}}$ encourages $\hat{z}^u_{3,j}$ to be similar to $z^u_{3,j}$ and different from z'^u:

$$\mathcal{L}_{\mathrm{CL}} = \sum_{j,u} \mathcal{L}_{j,u} \tag{10}$$

$$\mathcal{L}_{j,u} = -\log \frac{\exp\left(f(\hat{z}^u_{3,j}, z^u_{3,j})/\tau\right)}{\exp\left(f(\hat{z}^u_{3,j}, z^u_{3,j})/\tau\right) + \sum_{k=1}^{K} \exp\left(f(\hat{z}^u_{3,j}, z'^u_k)/\tau\right)} \tag{11}$$

where $f(a,b) = a^T \cdot b/(\|a\|\|b\|)$ is the cosine similarity function and τ denotes temperature. Then the total loss of our proposed model is $\mathcal{L}_{\mathrm{Total}} = \mathcal{L}_{\mathrm{CE}} + \mathcal{L}_{\mathrm{CL}}$.

3. EXPERIMENTS

3.1. Dataset

We train our model on Nottingham database [16] and the Chinese pop dataset [17]. Our dataset contains 2154 melodies (at song level) in total. We split these songs into 2 subsets: 95% songs for training and for 5% songs for testing. The training set is augmented by transposing the key of each song. We set $N = 4$ in our model. The data format is designed the same way as in [9]. We use a 130-dimensional one-hot vector to represent a melody token (128 dimensions for note onsets with 128 MIDI pitches, 1 for note sustain and 1 for rest), and a 3-dimensional one-hot vector for rhythm token (1 dimension for note onset of any pitch, 1 for sustain and 1 for rest). Each vector denotes a $\frac{1}{4}$-beat time step.

3.2. Training Details

The hidden dimension of the two-layer GRU in the bi-directional inpainting module is 1024 for both pitch and rhythm. The structure and hidden dimension of all range EC^2-VAEs are same, as shown in [13] and [9]. The parameter of each EC^2-VAE is as same as the model in the orginal paper [9, 13]. The latent dimension of disentangled representations from each range of EC^2-VAE encoder and fed into the decoder is 128. Our model is trained by Adam optimizer [18] with learning rate from 1e-3 to 1e-5. We adopt an early stopping strategy to prevent over-fitting. We set the batch size to 128, τ to 1 and the number of negative samples in K to 384.

3.3. Generated Examples

We show results of three representative songs in Fig. 3. For each example, we select 16 bars of the original song and let the model generate bars 9 to 12. For *Danny boy* and *Oh Susanna*, we compare the results of our model (shown as *ours*) with the baseline results, which are generated without contrastive loss and long-term representations (*b.l.*). Besides, for *Oh Susanna*, we add an extra experiment (*ours'*) in which we modify the last phrase and see how the inpainted result changes accordingly.

We also experiment on the model's ability to inpaint structured melody. The original *Frère Jacques* adopts the *ABAB* structure. The inpainted result (c1) shows a similar structure. In (c2), we changed the song structure to *AABB*, and our generative results closely follow the new structure. The overall experimental results demonstrate that our inpainting model can perform high-quality inpainting based on past and future contexts. We release more demos online[2].

3.4. Objective and Subjective Evaluations

To further check the individual impacts of long-term representation (*l.r.*) and contrastive loss (*c.l.*), we conducted both objective and subject studies. Table 1 shows the objective evaluation results. We see all models perform similarly on the

[2]https://sqwei17.github.io/inpaint_demo_page/.

Fig. 3: Generative results (in the orange rectangles).

	Total Recons.	Rhythm Recons.	Total NLL	Rhythm NLL
Proposed	0.7111	0.9281	**1.2919**	**0.2943**
Proposed w/o *l.r.*.	**0.7221**	**0.9287**	1.3872	0.3518
Proposed w/o *c.l.*	0.7044	0.9266	1.3622	0.3333
Baseline: w/o *c.l.*& *l.r.*	0.6951	0.9273	1.5345	0.3899

Table 1: Objective evaluation results of testing reconstruction accuracy and negative log-likelihood (NLL).

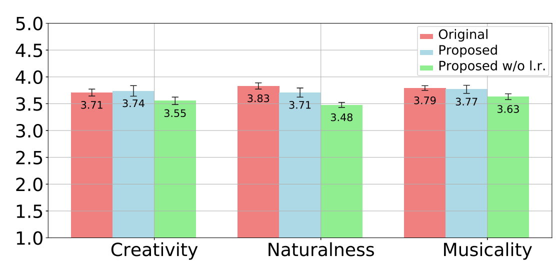

Fig. 4: Subjective evaluation results.

prediction accuracy, but our full model achieves significantly lower average NLL score compared to others.

Fig. 4 shows the subjective study where we further include original human composition as a version to compared with. 17 subjects (7 females and 10 males) participated in the study, during which each subject rates the results of different models in terms of *creativity*, *naturalness*, and overall *musicality*. The choice of the three indicators is explained in more detailed in [9]. In particular, each subject listen to 5 groups of samples, in which each group contains three generated versions: our proposed model, human compositions, and the proposed model without long-term representations.

We see that people prefer melodies generated by the proposed model to those generated without long-term representations. The performance of our model is even marginally better than human composition in terms of *creativity*. The heights of bars represent mean of the ratings and the error bars represent the MSEs computed via within-subject ANOVA [19].

4. CONCLUSION

In this paper, we propose a music inpainting model for phrase-level melody completion. We apply a forward-backward recurrent module for context-sensitive generation. The model performs prediction in a hierarchical manner to make use of both short-term and long-term context of past and future phrases. We also introduce the contrastive method to encourage the coherence of the generated phrase. Generative results show that our proposed model achieves better completion results compared to baseline. Our model can also inpaint structured melody that maintains consistency with the whole piece and the chord condition.

5. REFERENCES

[1] Adam Roberts, Jesse H. Engel, Colin Raffel, Curtis Hawthorne, and Douglas Eck, "A hierarchical latent vector model for learning long-term structure in music," in *Proceedings of the 35th International Conference on Machine Learning, ICML*, 2018, vol. 80, pp. 4361–4370.

[2] Cheng-Zhi Anna Huang, Ashish Vaswani, Jakob Uszkoreit, Ian Simon, Curtis Hawthorne, Noam Shazeer, Andrew M. Dai, Matthew D. Hoffman, Monica Dinculescu, and Douglas Eck, "Music transformer: Generating music with long-term structure," in *7th International Conference on Learning Representations, ICLR*, 2019.

[3] Ian Simon and Sageev Oore, "Performance rnn: Generating music with expressive timing and dynamics," *Magenta Blog*, 2017, `https://magenta.tensorflow.org/performance-rnn`.

[4] Gaëtan Hadjeres, François Pachet, and Frank Nielsen, "Deepbach: a steerable model for bach chorales generation," in *Proceedings of the 34th International Conference on Machine Learning, ICML*, Doina Precup and Yee Whye Teh, Eds., 2017, vol. 70, pp. 1362–1371.

[5] Gaëtan Hadjeres and Frank Nielsen, "Anticipation-rnn: enforcing unary constraints in sequence generation, with application to interactive music generation," *Neural Comput. Appl.*, vol. 32, no. 4, pp. 995–1005, 2020.

[6] Ashis Pati, Alexander Lerch, and Gaëtan Hadjeres, "Learning to traverse latent spaces for musical score inpainting," in *Proceedings of the 20th International Society for Music Information Retrieval Conference, ISMIR*, 2019, pp. 343–351.

[7] Ke Chen, Cheng-i Wang, Taylor Berg-Kirkpatrick, and Shlomo Dubnov, "Music sketchnet: Controllable music generation via factorized representations of pitch and rhythm," in *Proceedings of the 21th International Society for Music Information Retrieval Conference, ISMIR*, 2020, pp. 77–84.

[8] Cheng-Zhi Anna Huang, Tim Cooijmans, Adam Roberts, Aaron C. Courville, and Douglas Eck, "Counterpoint by convolution," in *Proceedings of the 18th International Society for Music Information Retrieval Conference, ISMIR*, 2017, pp. 211–218.

[9] Ruihan Yang, Dingsu Wang, Ziyu Wang, Tianyao Chen, Junyan Jiang, and Gus Xia, "Deep music analogy via latent representation disentanglement," in *Proceedings of the 20th International Society for Music Information Retrieval Conference, ISMIR*, 2019, pp. 596–603.

[10] Ziyu Wang, Dingsu Wang, Yixiao Zhang, and Gus Xia, "Learning interpretable representation for controllable polyphonic music generation," in *Proceedings of the 21th International Society for Music Information Retrieval Conference, ISMIR*, 2020, pp. 662–669.

[11] Junyan Jiang, Gus Xia, Dave B. Carlton, Chris N. Anderson, and Ryan H. Miyakawa, "Transformer VAE: A hierarchical model for structure-aware and interpretable music representation learning," in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*. 2020, pp. 516–520, IEEE.

[12] Hao-Wen Dong, Wen-Yi Hsiao, Li-Chia Yang, and Yi-Hsuan Yang, "Musegan: Multi-track sequential generative adversarial networks for symbolic music generation and accompaniment," in *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, 2018, pp. 34–41.

[13] Shiqi Wei and Gus Xia, "Learning long-term music representations via hierarchical contextual constraints," in *Proceedings of the 22nd International Society for Music Information Retrieval Conference, ISMIR*, 2021, pp. 343–351.

[14] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton, "A simple framework for contrastive learning of visual representations," in *Proceedings of the 37th International Conference on Machine Learning, ICML*, 2020, vol. 119 of *Proceedings of Machine Learning Research*, pp. 1597–1607.

[15] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick, "Momentum contrast for unsupervised visual representation learning," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition,CVPR*, 2020, pp. 9726–9735.

[16] E. Foxley, "Nottingham database," 2011, `http://abc.sourceforge.net/NMD/`.

[17] Ziyu Wang, Ke Chen, Junyan Jiang, Yiyi Zhang, Maoran Xu, Shuqi Dai, and Gus Xia, "POP909: A pop-song dataset for music arrangement generation," in *Proceedings of the 21th International Society for Music Information Retrieval Conference, ISMIR*, 2020, pp. 38–45.

[18] Diederik P. Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations, ICLR*, 2015.

[19] H Scheffé, "The analysis of variance," in *Architectural Institute of Japan*, 1999.