Pitch-aware generative pretraining improves multi-pitch estimation with scarce data

Mary Pilataki
Queen Mary University of London
London, United Kingdom
m.pilataki-manika@qmul.ac.uk

Matthias Mauch
Apple
London, United Kingdom
mmauch@apple.com

Simon Dixon
Queen Mary University of London
London, United Kingdom
s.e.dixon@qmul.ac.uk

Abstract

We demonstrate that pretrained generative models can learn representations that are useful for multi-pitch estimation. We explore representations extracted from DAC, a state-of-the-art audio compression model [24], which is based on VQ-GAN, an encoder-decoder architecture with vector quantisation. We propose pitch conditioning in the model's latent space such that the learned embeddings are pitch-aware. To determine whether such representations are suitable for transcription, we use them as input features to train a shallow multi-pitch transcriber. We show that conditioning the encoder with ground truth pitch targets leads to substantially improved transcription results. These improvements hold true even when conditioning on noisy labels generated by an off-the-shelf music transcriber, eliminating the need for annotated data during pretraining. Specifically, pitch conditioning in the pretraining phase yields an absolute average improvement of 14.5% and 12.0% in framewise and notewise F-scores respectively across datasets. Furthermore, we show that our representation learning method facilitates efficient transfer learning since our downstream model's performance is comparable to recent work even though it is trained on audio of a total duration of only 2 hours per dataset for 20 epochs. The source code of this work is available on Github [1].

CCS Concepts

• **Information systems** → **Music retrieval**; • **Computing methodologies** → *Artificial intelligence.*

Keywords

Music Transcription, Pitch Estimation, Music Information Retrieval, Audio Compression, Deep Learning

ACM Reference Format:
Mary Pilataki, Matthias Mauch, and Simon Dixon. 2024. Pitch-aware generative pretraining improves multi-pitch estimation with scarce data. In *ACM Multimedia Asia (MMASIA '24), December 3–6, 2024, Auckland, New Zealand*. ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3696409.3700202

[1] https://github.com/marypilataki/padac-mmasia24

MMASIA '24, December 3–6, 2024, Auckland, New Zealand
© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 979-8-4007-1273-9/24/12
https://doi.org/10.1145/3696409.3700202

1 Introduction

Multi-instrument music transcription is a fundamental task in the Music Information Retrieval (MIR) field. It is a challenging problem due to the complexity and variety of music signals, including issues such as overlapping harmonics and uncertainty regarding the number of simultaneously active notes or instruments. It consists of many subtasks, such as multi-pitch estimation (MPE), onset and offset time detection, instrument recognition, beat tracking, interpretation of expressive dynamics, and score typesetting [3].

Although deep learning has led to significant advances in multi-instrument transcription, the lack of annotated datasets and the increased hardware requirements impede research progress and practical applications of transcription systems. Furthermore, models are usually designed and trained in a task-specific manner, for example to target specified instruments, which means that they may generalize poorly or that they may be difficult to adapt for similar tasks.

Large pretrained models are commonly used in MIR either as feature extractors for downstream tasks [7, 13, 43] or for finetuning and transfer learning [33, 42]. Such approaches aim to eliminate the requirement for large amounts of annotated data, which is particularly useful in music transcription where labeled datasets are scarce due to the cost of manual annotation.

Descript Audio Codec (DAC) is a Vector Quantized Generative Adversarial Network (VQ-GAN) that demonstrates state of the art performance in audio compression. Motivated by 1) great progress in the generative audio field, 2) the fact that directly modeling music audio as opposed to time-frequency transformations or labels, yields richer representations for MIR [7] and 3) the introduction of reconstruction objectives in recent music transcription systems [8–10], we propose a VQ-GAN model architecture based on DAC [24], which performs music compression, reconstruction and pitch conditioning. We refer to this model as PA-DAC (Pitch-Aware DAC). We show that PA-DAC learns useful representations for multi-pitch estimation (MPE) and hence transcription.

Specifically, we propose pitch conditioning that guides the encoder module to learn embeddings that can describe pitch qualities. To evaluate whether our representation learning methodology is beneficial for music transcription, we pretrain PA-DAC on musical datasets and use it as a feature extractor to train a shallow framewise transcriber. Not only do results show that pitch awareness in the pretraining phase greatly benefits the downstream task of MPE, but they are also comparable to or outperform other SOTA transcription models.

Furthermore, we show that pitch conditioning is beneficial even if the ground truth pitch labels used for conditioning are noisy. To do this, we use an off-the shelf pretrained transcription model to synthetically generate pseudo-ground-truth for pitch conditioning which we refer to as *noisy labels*. Thus, we introduce flexibility in dataset choices in the pretraining phase as there is no requirement for annotated data. Using this method, the model can be pretrained on a diverse set of musical data (genre, instrument, recording conditions, etc.) which reduces the risk of overfitting and improves generalisation in downstream tasks.

This work makes the following contributions:

- We propose a two-stage methodology for representation learning and transfer learning for MPE. The first stage comprises of PA-DAC and the second stage of training a shallow transcriber for the downstream task of MPE using the representations learned in the first stage.
- We show that by introducing pitch awareness in the pretraining phase, the model used for the downstream task can be generic and while small amounts of data for training are sufficient.
- We propose pretraining on noisy labels using an off-the-shelf transcription model, thus eliminating the need for annotated data and allowing for flexibility in dataset choices.

The paper is structured as follows. In section 2 we present state of the art work on music transcription. We also discuss recent work on generative pretraining within the MIR field. In section 3 we describe the pretraining stage of our work: the PA-DAC model architecture and its training details as well as the noisy label generation method we used for supervised pitch conditioning. Section 4 describes the downstream model along with training and evaluation details. Section 5 presents and discusses the results. These are obtained from evaluating the downstream model which was trained on embeddings extracted from the pretrained PA-DAC. Section 6 concludes this paper.

2 Related Work

2.1 Music transcription

Deep learning models have become the current state of the art (SOTA) in Automatic Music Transcription (AMT). Many of these focus on transcription of specific instruments, mainly piano due to data availability [16, 21] as well as guitar [20, 31, 46], violin [2, 34], voice [22, 40, 41] and drums [6, 19] among others.

Multi-instrument AMT studies the problem of transcription of music signals that consist of multiple sources (different instruments and vocals). There are transcription systems that simultaneously estimate note events and their instrument sources [15, 18, 27, 28, 33, 44] and others that focus on estimating note events only [4, 8, 10, 11, 32, 45]. In the latter case, where instrument sources are not inferred, transcription systems are referred to as instrument-agnostic. Our work belongs to that category of system and we leave the instrument source estimation for future work.

SOTA transcription systems are commonly designed in a task-specific manner. They are based on either Recurrent Neural Networks (RNNs) such as Long Short-Term Memory networks [16, 21, 28], or U-Net structures [18], or Transformer models [15, 33, 36]. These architectures are popular for analyzing temporal sequences of acoustic features. In most cases, models are large hence resource-hungry, with MT3 for example reaching 60 million trainable parameters [15]. In the case of traditional supervised learning, large amounts of annotated data are required for training.

Addressing the problem of large and resource hungry models, Basic Pitch is the first lightweight (approximately 16,000 parameters) model for transcription [4]. It is an extension of Deep Salience, a fully convolutional instrument agnostic model with a Harmonic Constant-Q Transform (HCQT) frontend [5]. It consists of a multi-output structure that estimates framewise onsets, multipitch and note activations and is followed by a post-processing mechanism to produce note-level estimates. Despite its simplicity and low-resource setting, the model can achieve comparable results to other SOTA models.

To counteract the requirement for large annotated datasets, there are a few approaches that introduce semi-supervised and continual learning [8] and self-supervised learning [11, 32]. The aforementioned models are instrument-agnostic. ReconVAT [8] uses Virtual Adversarial Training (VAT) which enables semi-supervised learning with unlabelled data in combination with small annotated datasets. PESTO [32] leverages equivariance to musical transpositions and proposes a Siamese-style network that learns to capture pitch information given pairs of transposed inputs. SS-MPE [11] uses an autoencoder model inspired by Timbre Trap [10] together with self-supervised objectives which leverage the properties of timbre invariance and geometric equivariance.

Several transcription models have shown that incorporating reconstruction objectives in their system is beneficial for transcription. ReconVAT [8, 9] includes a reconstruction module within its transcription system which learns to reconstruct the input spectrogram using the posteriogram produced by a first-step transcription process. The reconstructed spectrogram is then used to produce the transcription output. It is shown that the reconstructed spectrogram is a denoised version of the original spectrogram and the transcription produced using the reconstruction yields improved performance.

Timbre Trap [10] is an autoencoder model that can operate in two modes: MPE and reconstruction of complex CQT coefficients. In MPE mode, the decoder outputs pitch salience probabilities, while in reconstruction mode, the decoder outputs real and imaginary parts of CQT coefficients. The model architecture is based on audio compression models [14, 47] and hence is similar to the model we use in the pretraining phase.

Timbre Trap is the first work to introduce the idea of a unified model for transcription and reconstruction with promising results. Building on top of this idea, we propose a two-stage methodology instead, where the first stage involves representation and reconstruction learning and the second stage involves transcription and transfer learning to new datasets.

2.2 Generative pretraining for MIR

Previous research has shown that using pretrained generative models for discriminative MIR tasks is beneficial. JukeMIR [7] and Sheet Sage [13] utilise representations from Jukebox [12], a VQ-VAE music generation model which contains a language model trained on codified audio. The suitability of representations is evaluated for

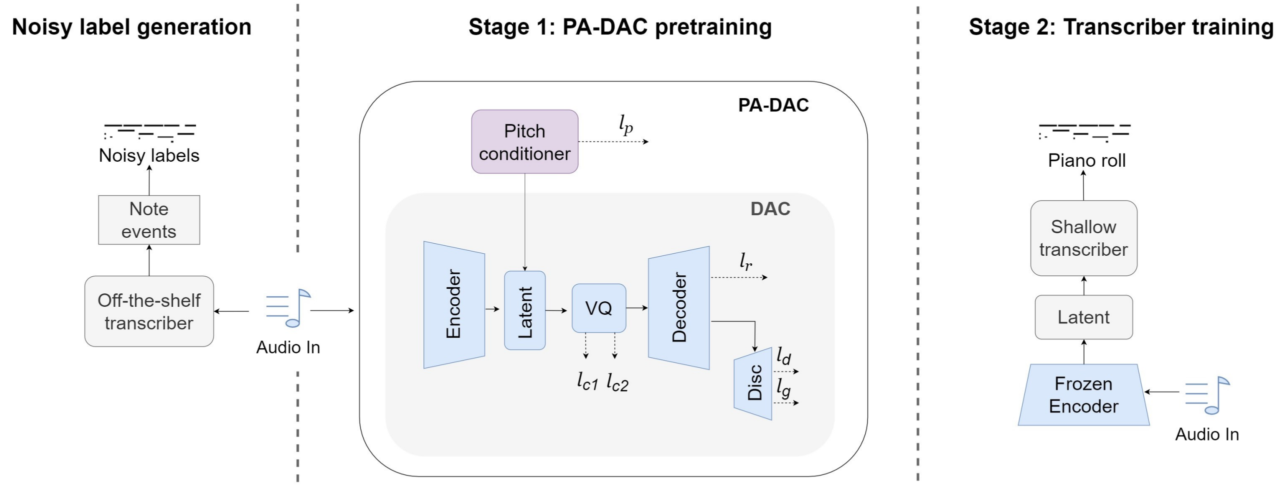


Figure 1: Noisy labels (left) are generated by converting note events estimated by an off-the-shelf transcriber into piano rolls. These are used for pitch conditioning within PA-DAC. PA-DAC (middle) produces pitch-aware latent space embeddings via a pitch conditioning module. The shaded part of the diagram corresponds to the original DAC model. Dotted arrows show the losses associated with each module. l_p denotes the pitch loss, l_{c1} and l_{c2} the codebook and commitment losses, l_r the reconstruction loss, l_d and l_g the discriminator and generator losses. After pretraining PA-DAC, latent space embeddings are extracted from the encoder to train a shallow transcriber (right).

several downstream MIR tasks: tagging, genre classification, key detection, emotion recognition [7] and melody transcription [13]. It is shown that generative pretraining greatly improves performance on the downstream tasks when compared to handcrafted features such as time-frequency transformations or features extracted from models trained to perform other discriminative tasks. Both JukeMIR and Sheet Sage achieve comparable performance to SOTA even though JukeMIR uses a generic shallow model for all downstream tasks.

While our work can be considered similar to JukeMIR [7] and Sheet Sage [13], our proposed method differs in two ways: 1) we further repurpose the generative model we use in the pretraining phase such that representation learning is specific to our downstream task; and 2) we choose the embeddings learned from the encoder module to train our downstream model while the previous methods [7, 13] use embeddings learned from the language model within Jukebox which is trained on vector-quantised tokens.

3 PA-DAC Pretraining

3.1 Model architecture

PA-DAC is a VQ-GAN model based on DAC [24]. It consists of a fully convolutional encoder-decoder structure which performs temporal downscaling with a chosen stride factor, a Residual Vector Quantisation (RVQ) module which compresses the latent space into discrete tokens before entering the decoding stage and a complex STFT discriminator. The model accepts variable-length audio sampled at 44.1 kHz. For further details regarding the aforementioned modules please refer to the original DAC paper [24].

PA-DAC incorporates a pitch conditioning module which drives the encoder to produce embeddings that can describe pitch. The architecture is shown in Figure 1 (middle). The pitch conditioning module is a two-layer fully connected network (FCN) with each layer followed by a ReLU activation function. The pitch conditioner accepts latent space embeddings of one second of audio. These are of dimensionality $T \times D = 87 \times 1024$, where T denotes time and D the latent space size. It outputs 128 features for each one of the 87 frames which represent the probability of each pitch being active at that frame.

3.2 Training objectives

Following a multitask learning configuration, the PA-DAC model optimises several objectives simultaneously which are also depicted in Figure 1 (middle):

frequency domain reconstruction with a multi-scale L1 loss for mel spectrograms

time and frequency domain discrimination using the Hinge-GAN adversarial loss formulation [25] and the L1 feature matching loss [23]

codebook learning with the original codebook and commitment losses from the VQ-VAE formulation [38]

pitch conditioning with the binary cross entropy loss.

To balance the influence of each loss term, we apply a different weight to each. For the reconstruction, discrimination and codebook learning objectives, we adopt the same weights used in the original DAC model [24]. For the pitch conditioning term, we experimented with several weight values from 15.0 to 300.0 and we made the following observations: 1) the pitch conditioner weight greatly impacts the embeddings produced by the encoder hence it also impacts the performance of our transcriber probe and 2) the greater the weight, the greater the performance of our transcriber up to a certain threshold, when transcription performance starts to degrade again. We speculate that this might be due to mistakes in the noisy labels being weighted more strongly thus pitch conditioning becoming less accurate. For our final experiment, we choose the a weight of 150.0 for the pitch loss that yields embeddings which lead to the best performance for our transcriber. The overall loss is shown in Equation 1, where L_{recon} denotes the reconstruction loss, L_{adv} the adversarial loss, L_{feat} the feature matching loss, L_{code} the codebook loss, L_{commit} the commitment loss and L_{pitch} the pitch conditioning loss.

$$\begin{aligned} L_{\text{total}} = 15.0 \times L_{\text{recon}} + 1.0 \times L_{\text{adv}} + 2.0 \times L_{\text{feat}} \\ + 1.0 \times L_{\text{code}} + 0.25 \times L_{\text{commit}} + 150.0 \times L_{\text{pitch}} \end{aligned} \tag{1}$$

3.3 Noisy label generation

The pitch conditioning module within PA-DAC is trained in a supervised manner. We generate synthetic pitch labels using the pretrained Basic Pitch model [4] provided by its authors in the associated repository[2]. As shown in Figure 1 (left), we convert the estimated note events into piano rolls and use those as labels to train PA-DAC. We choose Basic Pitch because it is a lightweight model with a short inference time, it is trained on a variety of datasets including different instruments and it achieves comparable to SOTA perfomance, as described in section 2.1. For further details on the model, please refer to the original paper and the associated repository [4].

3.4 Pretraining details

We train PA-DAC using the following datasets:

Mazurkas [3] A collection of mazurkas of a total of 123 hours performed by 157 different pianists.

Bach violin dataset Introduced by Tamer et al. [34], this is a 34-hour dataset of solo violin recordings composed by Bach.

Guitar dataset A collection of audio-score pairs that have been used for training deep learning models for guitar transcription [31].

GTZAN A collection of 100 30-second excerpts for each of the following music genres: blues, classical, country, disco, hip-hop, jazz, metal, pop, reggae, rock [37]. For our experiments, we only use the blues, classical, country, jazz, pop and rock subsets.

We train the model for 100k iterations with a batch size of 6. At each iteration, the model accepts a 1-second audio excerpt from each one of the datasets except for GTZAN from which the model sees three examples from each one of the following subsets: 1) classical, 2) jazz and 3) blues, country, pop or rock.

Following the original DAC paper [24], we use the AdamW optimizer [26] with a learning rate of 0.0001, $\beta_1 = 0.8$ and $\beta_2 = 0.8$ for both the generator and the discriminator. We reduce the learning rate at every step with $\gamma = 0.999996$.

3.5 Ablation experiment

In order to evaluate the quality of the synthetic labels generated using the method described in section 3.3, we also run an ablation experiment, where we pretrain PA-DAC using the MAESTRO dataset [17] in two different conditions: 1) using the annotations provided with the dataset as ground truth; and 2) using Basic Pitch predictions as ground truth. All training details and objectives are identical to those described in sections 3.4 and 3.2 except for the dataset; for the ablation experiment we only use MAESTRO for pretraining PA-DAC. We then train our transcriber probe using features extracted from each condition and compare the results.

4 Probing method

4.1 Model architecture

To evaluate the effectiveness of our proposed representation learning method, we use a shallow classifier, referred to as a 'probe'. A probe can only use the hidden units of a given intermediate layer as discriminating features. This method was introduced by Alain and Bengio [1] and adopted by Castellon et al. [7] for probing Jukebox representations, as described in section 2.2.

To this end, we train a supervised one-layer perceptron with 512 hidden units. The model accepts as inputs latent space embeddings of 1 second of audio which are learned by PA-DAC. Each input has a dimensionality of $T \times D = 87 \times 1024$ where T denotes time and D denotes the latent space dimensionality. The model is trained to predict probabilities of each pitch being active at each point in time. To achieve this, model outputs are passed through a sigmoid activation function. Those outputs are of dimensionality $T \times P = 87 \times 128$. P denotes MIDI pitches 0-127, based on the equal-tempered scale [4].

4.2 Training details

After pretraining PA-DAC, its weights are frozen and the latent space embeddings are used as inputs to our transcriber probe. We use the PA-DAC checkpoint from iteration 100k. We first split audio into 1 second excerpts, 0.2 seconds overlapping and feed those into PA-DAC. We extract and save the features. The corresponding diagram is shown in Figure 1 (right).

We then train the model in a supervised manner on subsets of Slakh [29], MusicNet [35] and GuitarSet [46]. We randomly sample audio excerpts of a total duration of two hours per dataset. For validation we use random excerpts of a total duration of 10% that of our training data, 12 minutes per dataset. We exclude tracks used for training. We provide the exact track IDs for each split in our open-source code release for this paper. We train the model for 20 epochs with a batch size of 2. The Adam optimiser is used with a learning rate of 0.00005 and a weight decay of 0.00001.

4.3 Evaluation details

For evaluating, we use the checkpoint of the model at epoch number 20. We test the model on the official test sets of Slakh and MusicNet and the full GuitarSet dataset. We also include a comparison with the following instrument-agnostic models where results are available: Timbre Trap [10], SS-MPE [11], Deep Salience [5] and Basic Pitch [4]. We also include a comparison with MT3, a SOTA multi-instrument transcription model [15]. The experimental setup differs between the aforementioned baselines (model architectures, training details and datasets are highly different), hence direct comparisons cannot be made.

[2] https://github.com/spotify/basic-pitch/

[3] http://www.charm.rhul.ac.uk/index.html

[4] https://midi.org/

Table 1: Precision (P), recall (R), F-score (F) and accuracy (Acc) percentages of the transcriber probe MPE estimates and note-level F-scores (Fn). The leftmost column indicates the pretrained model used to extract features to train the transcriber. PA-DAC (n) and PA-DAC (g) refer to the ablation experiment discussed in Section 3.5 where pretraining is performed on MAESTRO only using the dataset's ground truth (g) or noisy labels (n).

	GuitarSet					*Slakh*					*MusicNet*				
	P	*R*	*F*	*Acc*	*Fn*	*P*	*R*	*F*	*Acc*	*Fn*	*P*	*R*	*F*	*Acc*	*Fn*
Pretrained model															
DAC	64.7	64.0	64.4	43.3	37.9	62.8	55.5	58.9	40.9	30.8	51.9	44.0	47.6	28.6	24.3
PA-DAC	84.6	77.0	**80.6**	66.7	**49.4**	66.7	73.0	**69.7**	52.7	42.9	59.5	69.1	**64.0**	45.5	**36.8**
PA-DAC (g)	79.1	68.5	73.4	56.3	43.5	66.4	67.6	67.0	49.6	37.7	56.7	63.7	60.0	41.3	33.5
PA-DAC (n)	77.4	68.7	72.8	55.3	42.5	67.4	65.6	66.5	48.9	37.1	59.8	57.8	58.8	39.6	32.1

Table 2: Baseline comparison of frame-level Precision (P), Recall (R), F-score (F) and Accuracy (Acc), plus note-level F-score (Fn) (where applicable). The leftmost column indicates the transcription model except for our method which indicates the model used in the pretraining phase. The asterisk indicates that GuitarSet is excluded from the training data.

	GuitarSet					*Slakh*					*MusicNet*				
	P	*R*	*F*	*Acc*	*Fn*	*P*	*R*	*F*	*Acc*	*Fn*	*P*	*R*	*F*	*Acc*	*Fn*
Method															
Timbre Trap	48.6	75.6	59.2	41.3	-	-	-	-	-	-	-	-	-	-	-
SS-MPE	59.5	58.1	58.8	40.2	-	-	-	-	-	-	-	-	-	-	-
Deep Salience	77.7	70.6	74.0	57.7	-	-	-	-	-	-	-	-	-	-	-
MT3	-	-	-	-	-	-	-	**79.0**	-	**76.0**	-	-	**68.0**	-	50.0
Basic Pitch	79.4	85.1	**82.2**	69.5	**77.6**	72.9	39.4	51.2	34.0	40.5	65.1	51.2	57.3	39.8	**62.3**
PA-DAC	84.6	77.0	80.6	66.7	49.4	66.7	73.0	69.7	52.7	42.9	59.5	69.1	64.0	45.5	36.8
PA-DAC*	63.6	77.8	70.0	52.2	35.7	66.0	74.0	69.8	52.8	43.6	57.2	69.6	62.8	44.3	35.5

We utilize the community-standard mir_eval package [30] to compute precision (P), recall (R), F-score (F) and accuracy (Acc) of multi-pitch estimates with respect to the ground truth. Frame-level P-R-F estimates are obtained by performing local peak-picking across frequency and applying a threshold of 0.3. Accuracy is computed as the ratio between the number of pitch estimates within 0.5 semitones of matched ground-truth pitches to the total number of pitch estimates plus the number of missed ground-truth pitch estimates.

To compute note-level metrics, we convert frame estimates to note events by applying the post-processing method introduced by Bittner et al. [4]. The only modification we make is that we estimate onset times using our model's frame output. In that case, a threshold of 0.1 is used and detected events that are shorter than 10 frames (≈ 116 ms) are removed. For further details on the post-processing method please refer to the original Basic Pitch paper [4]. Notes are considered correct if the pitch is within a quarter tone and the onset is within 50 ms. Final framewise and notewise results are computed by averaging across all tracks within an individual dataset.

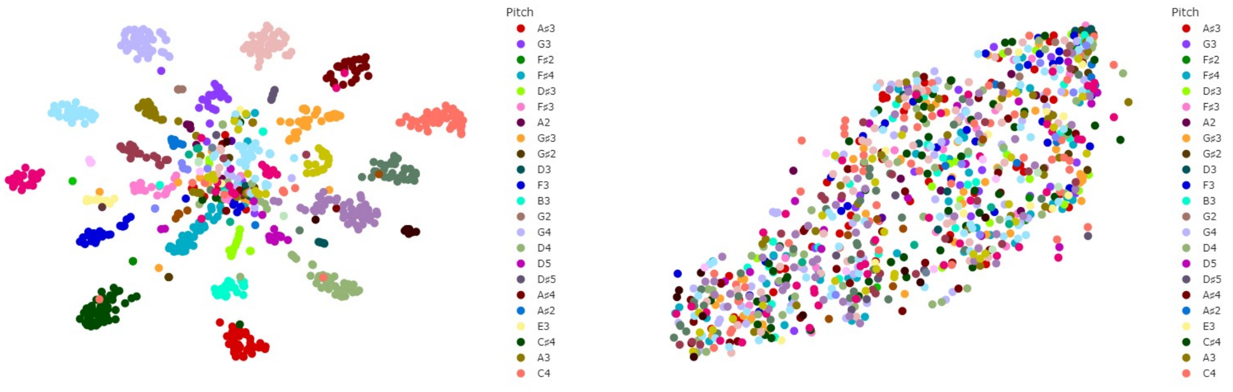
5 Results

Table 1 presents the performance of the multi-pitch transcriber probe trained on representations extracted from the pretrained DAC and PA-DAC models. Pretraining is done using the datasets described in section 3 except for PA-DAC (g) and PA-DAC (n) (last two rows in the table) which are pretrained on the MAESTRO dataset only. PA-DAC (g) and PA-DAC (n) correspond to results of our ablation experiment which is described in section 3.5. (n) denotes that the model is trained on noisy labels whereas (g) denotes that the model is trained on MAESTRO's ground truth annotations.

For PA-DAC, the noisy labels for pitch conditioning are generated using the method described in section 3.3. It is apparent that pitch conditioning in the pretraining phase not only greatly benefits multi-pitch estimation, but also the estimation of note events after post-processing. Specifically, the percentage increase for each of the datasets, GuitarSet, Slakh and MusicNet respectively, is as follows: 1) Frame accuracy: 23.4, 11.8 and 16.9; 2) Frame F-score: 16.2, 10.8 and 16.4; 3) Note F-score: 11.5, 12.1 and 12.5.

Referring to Table 1 and comparing PA-DAC (g) and PA-DAC (n), results indicate that using noisy labels in the pretraining phase does not significantly degrade the downstream task performance. Specifically, when using noisy labels in pretraining there is an average performance decrease of 0.8%, 1% and 1.1% in frame-wise and note-wise f-scores and frame-wise accuracy respectively across datasets compared to when using the ground truth. Although pitch conditioning is more accurate when using ground truth, the value of pitch conditioning on noisy labels is apparent. The performance increase in pitch estimation when using representations extracted from PA-DAC (n) compared to those extracted from DAC is 9.1%, 6.2% and 10.3% for frame-wise and note-wise f-scores and frame-wise accuracy respectively across datasets.

As expected, representations extracted from PA-DAC yield a stronger performance compared to representations extracted from PA-DAC (g) and PA-DAC (n). PA-DAC is trained on several datasets, described in section 3, hence the learned representations are richer, capturing timbre and pitch related information for a variety of instruments. Nevertheless, with PA-DAC (g) and PA-DAC (n), transfer



(a) PA-DAC model's latent space.

(b) DAC model's latent space.

Figure 2: t-SNE visualization of the latent space extracted from PA-DAC (a) and DAC (b) using the GuitarSet dataset. A subset of 20 pitch classes are shown. Colours represent different pitch classes.

learning to instruments other than piano is effective as can be seen in Table 1.

Table 2 presents the performance of several instrument-agnostic models along with our best model which is trained on representations extracted from PA-DAC. We use the checkpoint provided online by the authors of Basic Pitch [4] to evaluate their model. We obtain results for SS-MPE [11], Timbre Trap [10] and Deep Salience [5] from the SS-MPE paper [11]. In that case, the evaluation is performed using the whole GuitarSet dataset as this was not included in the training data for the aforementioned models. For a fairer comparison, we also re-train our model excluding all GuitarSet tracks from training and validation. This model is denoted in Table 2 as PA-DAC*. Finally, we obtain MT3 results from the original paper [15].

Our method outperforms Timbre Trap [10] and SS-MPE [11] across all metrics while performance is comparable to Deep Salience [5]. Furthermore, PA-DAC* performs well on the GuitarSet dataset even though it is only trained on a few synthetic guitar examples from Slakh.

Our method clearly outperforms Basic Pitch in terms of framewise performance on Slakh and MusicNet, while for GuitarSet framewise scores are comparable. In terms of notewise performance, our method achieves a better score on Slakh while for the other two datasets Basic Pitch performs better by a large margin. This is expected because Basic Pitch is a specialised architecture which combines framewise onsets, multipitch and note activations, to predict note events in contrast to our model which only uses frame estimates. Furthermore, Basic Pitch is trained on the Guitarset dataset while PA-DAC* excludes all Guitarset data from its training set.

Finally, as expected due to its sophisticated architecture and parameter size, MT3 is the best performing model overall and outperforms our model by a large margin. However, MT3 is outperformed by Basic Pitch on the MusicNet test set in terms of note-wise f-score.

We visualise the latent space of our pretrained models using the t-SNE dimensionality reduction method [39]. Figure 2 (a) and (b) shows the latent space of PA-DAC and DAC respectively where different colours indicate different pitch classes. For clearer visualisation, those embeddings have been extracted using monophonic excerpts from the GuitarSet dataset. The exact track IDs are included in the open-source code release for this paper. We observe that pitch conditioning introduces the formation of pitch clusters as opposed to the original DAC model.

6 Conclusion

In this paper, we have described a representation learning method based on pretraining a VQ-GAN model. We have introduced pitch-awareness in the pretraining phase and have shown that this greatly aids the downstream task of multi-pitch estimation and facilitates efficient transfer learning with scarce data. Our downstream model was trained on only 2 hours of data per dataset for 20 epochs, proving that our method is suitable for low-resource settings. We have also proposed pretraining on noisy labels where there is no requirement for labeled data, allowing for flexibility in dataset choices. Although in this work we use Basic Pitch [4] to synthetically generate noisy labels, different transcription models could be tested.

For future work we would also like to explore modifications in the pitch conditioning module such that conditioning is performed at the note-level, and replace our downstream model such that we can achieve increases in both frame and note-level scores for transcription. Finally, we are interested in evaluating the reconstruction capabilities of PA-DAC and exploring whether pitch conditioning also benefits music reconstruction .

Acknowledgments

This work is supported by the EPSRC UKRI Centre for Doctoral Training in Artificial Intelligence and Music (Grant no. EP/S022694/1).

References

[1] Guillaume Alain and Yoshua Bengio. 2016. Understanding intermediate layers using linear classifier probes. In *International Conference on Learning Representations (ICLR)*.

[2] Kemal Avci and Tamer Şevki Acuner. 2017. Automatic transcription of open string notes from violin recordings. In *25th Signal Processing and Communications Applications Conference (SIU)*. 1–4. https://doi.org/10.1109/SIU.2017.7960729

[3] Emmanouil Benetos, Simon Dixon, Zhiyao Duan, and Sebastian Ewert. 2019. Automatic Music Transcription: An Overview. *IEEE Signal Processing Magazine* 36, 1 (2019), 20–30. https://doi.org/10.1109/MSP.2018.2869928

[4] Rachel M. Bittner, Juan José Bosch, David Rubinstein, Gabriel Meseguer-Brocal, and Sebastian Ewert. 2022. A Lightweight Instrument-Agnostic Model for Polyphonic Note Transcription and Multipitch Estimation. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 781–785. https://doi.org/10.1109/ICASSP43922.2022.9746549

[5] Rachel M. Bittner, Brian McFee, Justin Salamon, Peter Qi Li, and Juan Pablo Bello. 2017. Deep Salience Representations for F0 Estimation in Polyphonic Music. In *International Society for Music Information Retrieval Conference (ISMIR)*. https://api.semanticscholar.org/CorpusID:4531539

[6] Lee Friese Callender, Curtis Glenn-Macway Hawthorne, and Jesse Engel. 2020. Improving Perceptual Quality of Drum Transcription with the Expanded Groove MIDI Dataset. *ArXiv* (2020). https://arxiv.org/abs/2004.00188

[7] Rodrigo Castellon, Chris Donahue, and Percy Liang. 2021. Codified audio language modeling learns useful representations for music information retrieval. In *International Society for Music Information Retrieval Conference (ISMIR)*.

[8] Kin Wai Cheuk, Dorien Herremans, and Li Su. 2021. ReconVAT: A Semi-Supervised Automatic Music Transcription Framework for Low-Resource Real-World Data. In *Proceedings of the 29th ACM International Conference on Multimedia* (Virtual Event, China) *(MM '21)*. Association for Computing Machinery, New York, NY, USA, 3918–3926. https://doi.org/10.1145/3474085.3475405

[9] Kin Wai Cheuk, Yin-Jyun Luo, Emmanouil Benetos, and Dorien Herremans. 2021. The Effect of Spectrogram Reconstruction on Automatic Music Transcription: An Alternative Approach to Improve Transcription Accuracy. In *25th International Conference on Pattern Recognition (ICPR)*. 9091–9098. https://doi.org/10.1109/ICPR48806.2021.9412155

[10] Frank Cwitkowitz, Kin Wai Cheuk, Woosung Choi, Marco A. Martínez-Ramírez, Keisuke Toyama, Wei-Hsiang Liao, and Yuki Mitsufuji. 2024. Timbre-Trap: A Low-Resource Framework for Instrument-Agnostic Music Transcription. In *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Seoul, Korea.

[11] Frank Cwitkowitz and Zhiyao Duan. 2024. Toward Fully Self-Supervised Multi-Pitch Estimation. *arXiv preprint arXiv:2402.15569* (2024).

[12] Prafulla Dhariwal, Heewoo Jun, Christine Payne, Jong Wook Kim, Alec Radford, and Ilya Sutskever. 2020. Jukebox: A Generative Model for Music. *arXiv preprint arXiv:2005.00341* (2020).

[13] Chris Donahue, John Thickstun, and Percy Liang. 2022. Melody transcription via generative pre-training. In *International Society for Music Information Retrieval Conference (ISMIR)*.

[14] Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. 2022. High Fidelity Neural Audio Compression. *arXiv preprint arXiv:2210.13438* (2022).

[15] Josh Gardner, Ian Simon, Ethan Manilow, Curtis Hawthorne, and Jesse Engel. 2021. MT3: Multi-Task Multitrack Music Transcription. In *International Conference on Learning Representations (ICLR)*.

[16] Curtis Hawthorne, Erich Elsen, Jialin Song, Adam Roberts, Ian Simon, Colin Raffel, Jesse Engel, Sageev Oore, and Douglas Eck. 2018. Onsets and Frames: Dual-Objective Piano Transcription. In *International Society for Music Information Retrieval Conference (ISMIR)*.

[17] Curtis Hawthorne, Andriy Stasyuk, Adam Roberts, Ian Simon, Cheng-Zhi Anna Huang, Sander Dieleman, Erich Elsen, Jesse Engel, and Douglas Eck. 2019. Enabling Factorized Piano Music Modeling and Generation with the MAESTRO Dataset. In *International Conference on Learning Representations (ICML)*. https://openreview.net/forum?id=r1lYRjC9F7

[18] Yun-Ning Hung, Yi-An Chen, and Yi-Hsuan Yang. 2019. Multitask Learning for Frame-level Instrument Recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 381–385. https://doi.org/10.1109/ICASSP.2019.8683426

[19] Daichi Kamakura, Eita Nakamura, and Kazuyoshi Yoshii. 2023. CTC2: End-to-End Drum Transcription Based on Connectionist Temporal Classification With Constant Tempo Constraint. In *Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. 158–164. https://doi.org/10.1109/APSIPAASC58517.2023.10317515

[20] Sehun Kim, Tomoki Hayashi, and Tomoki Toda. 2022. Note-level Automatic Guitar Transcription Using Attention Mechanism. In *30th European Signal Processing Conference (EUSIPCO)*. 229–233. https://doi.org/10.23919/EUSIPCO55093.2022.9909659

[21] Qiuqiang Kong, Bochen Li, Xuchen Song, Yuan Wan, and Yuxuan Wang. 2020. High-Resolution Piano Transcription With Pedals by Regressing Onset and Offset Times. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29 (2020), 3707–3717. https://api.semanticscholar.org/CorpusID:222133261

[22] Sangeun Kum, Jongpil Lee, Keunhyoung Luke Kim, Taehyoung Kim, and Juhan Nam. 2022. Pseudo-Label Transfer from Frame-Level to Note-Level in a Teacher-Student Framework for Singing Transcription from Polyphonic Music. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 796–800. https://doi.org/10.1109/ICASSP43922.2022.9747147

[23] Kundan Kumar, Rithesh Kumar, Thibault de Boissiere, Lucas Gestin, Wei Zhen Teoh, Jose Sotelo, Alexandre de Brebisson, Yoshua Bengio, and Aaron Courville. 2019. MelGAN: generative adversarial networks for conditional waveform synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2019/file/6804c9bca0a615bdb9374d00a9fcba59-Paper.pdf

[24] Rithesh Kumar, Prem Seetharaman, Alejandro Luebs, Ishaan Kumar, and Kundan Kumar. 2023. High-Fidelity Audio Compression with Improved RVQGAN. In *Advances in Neural Information Processing Systems (NeurIPS)*, A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (Eds.), Vol. 36. Curran Associates, Inc., 27980–27993. https://proceedings.neurips.cc/paper_files/paper/2023/file/58d0e78cf042af5876e12661087bea12-Paper-Conference.pdf

[25] Jae Hyun Lim and J. C. Ye. 2017. Geometric GAN. *ArXiv* abs/1705.02894 (2017). https://api.semanticscholar.org/CorpusID:9010805

[26] Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations (ICLR)*.

[27] Ben Maman and Amit H Bermano. 2022. Unaligned Supervision for Automatic Music Transcription in The Wild. In *Proceedings of the 39th International Conference on Machine Learning (ICML) (Proceedings of Machine Learning Research, Vol. 162)*, Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (Eds.). PMLR, 14918–14934. https://proceedings.mlr.press/v162/maman22a.html

[28] Ethan Manilow, Prem Seetharaman, and Bryan Pardo. 2020. Simultaneous Separation and Transcription of Mixtures with Multiple Polyphonic and Percussive Instruments. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 771–775. https://doi.org/10.1109/ICASSP40776.2020.9054340

[29] Ethan Manilow, Gordon Wichern, Prem Seetharaman, and Jonathan Le Roux. 2019. Cutting Music Source Separation Some Slakh: A Dataset to Study the Impact of Training Data Quality and Quantity. *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)* (October 2019). https://doi.org/10.1109/waspaa.2019.8937170

[30] Colin Raffel, Brian McFee, Eric J. Humphrey, Justin Salamon, Oriol Nieto, Dawen Liang, and Daniel P. W. Ellis. 2014. mir eval: A Transparent Implementation of Common MIR Metrics. In *International Society for Music Information Retrieval Conference (ISMIR)*, Hsin-Min Wang, Yi-Hsuan Yang, and Jin Ha Lee (Eds.). 367–372. http://dblp.uni-trier.de/db/conf/ismir/ismir2014.html#RaffelMHSNLE14

[31] Xavier Riley, Drew Edwards, and Simon Dixon. 2024. High Resolution Guitar Transcription via Domain Adaptation. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Seoul, Korea.

[32] Alain Riou, Stefan Lattner, Gaëtan Hadjeres, and Geoffroy Peeters. 2023. PESTO: Pitch Estimation with Self-supervised Transposition-equivariant Objective. In *Proceedings of the 24th Conf. of the International Society for Music Information Retrieval (ISMIR)*. Milan, Italy.

[33] Ian Simon, Joshua Gardner, Curtis Hawthorne, Ethan Manilow, and Jesse Engel. 2022. Scaling Polyphonic Transcription with Mixtures of Monophonic Transcriptions. In *Proceedings of the 23rd International Society for Music Information Retrieval Conference (ISMIR)*. ISMIR, Bengaluru, India, 44–51. https://doi.org/10.5281/zenodo.7316590

[34] Nazif Can Tamer, Yigitcan Özer, Meinard Müller, and Xavier Serra. 2023. High Resolution Violin Transcription using weak labels. In *International Society for Music Information Retrieval Conference (ISMIR)*.

[35] John Thickstun, Zaid Harchaoui, and Sham Kakade. 2017. Learning Features of Music from Scratch. In *International Conference on Learning Representations (ICLR)*.

[36] Keisuke Toyama, Taketo Akama, Yukara Ikemiya, Yuhta Takida, Wei-Hsiang Liao, and Yuki Mitsufuji. 2023. Automatic Piano Transcription with Hierarchical Frequency-Time Transformer. In *Proceedings of the 24th International Society for Music Information Retrieval Conference*.

[37] G. Tzanetakis and P. Cook. 2002. Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing* 10, 5 (2002), 293–302. https://doi.org/10.1109/TSA.2002.800560

[38] Aaron van den Oord, Oriol Vinyals, and koray kavukcuoglu. 2017. Neural Discrete Representation Learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/file/7a98af17e63a0ac09ce2e96d03992fbc-Paper.pdf

[39] Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing Data using t-SNE. *Journal of Machine Learning Research* 9, 86 (2008), 2579–2605. http://jmlr.org/papers/v9/vandermaaten08a.html

[40] Jun-You Wang and Jyh-Shing Roger Jang. 2021. On the Preparation and Validation of a Large-Scale Dataset of Singing Transcription. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 276–280. https://doi.org/10.1109/ICASSP39728.2021.9414601

[41] Jun-You Wang and Jyh-Shing Roger Jang. 2023. Training a Singing Transcription Model Using Connectionist Temporal Classification Loss and Cross-Entropy Loss. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 31 (2023), 383–396. https://doi.org/10.1109/TASLP.2022.3224297

[42] Ziyu Wang and Gus Xia. 2021. MuseBERT: Pre-training Music Representation for Music Understanding and Controllable Generation. In *Proceedings of the 22nd International Society for Music Information Retrieval Conference (ISMIR)*. ISMIR, Online, 722–729. https://doi.org/10.5281/zenodo.5624387

[43] Ho-Hsiang Wu, Chieh-Chi Kao, Qingming Tang, Ming Sun, Brian McFee, Juan Pablo Bello, and Chao Wang. 2021. Multi-Task Self-Supervised Pre-Training for Music Classification. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2021), 556–560. https://api.semanticscholar.org/CorpusID:231839503

[44] Yu-Te Wu, Berlin Chen, and Li Su. 2019. Polyphonic Music Transcription with Semantic Segmentation. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 166–170. https://doi.org/10.1109/ICASSP.2019.8682605

[45] Y. T. Wu, B. Chen, and L. Su. 2020. Multi-Instrument Automatic Music Transcription With Self-Attention-Based Instance Segmentation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 28 (2020), 2796–2809. https://doi.org/10.1109/TASLP.2020.3030482

[46] Qingyang Xi, Rachel M. Bittner, Johan Pauwels, Xuzhou Ye, and Juan Pablo Bello. 2018. GuitarSet: A Dataset for Guitar Transcription. In *International Society for Music Information Retrieval Conference (ISMIR)*.

[47] Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi. 2022. SoundStream: An End-to-End Neural Audio Codec. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 30 (2022), 495–507. https://doi.org/10.1109/TASLP.2021.3129994

Received 08 August 2024; accepted 04 October 2024; revised 18 October 2024