Multi-Class-Token Transformer for Multitask Self-supervised Music Information Retrieval

Yuexuan Kong, Vincent Lostanlen, Romain Hennequin, Mathieu Lagrange, Gabriel Meseguer-Brocal

Multi-Class-Token Transformer for Multitask Self-supervised Music Information Retrieval

Yuexuan Kong[1,2]*, Vincent Lostanlen*[2]*, Romain Hennequin*[1]*, Mathieu Lagrange*[2]*, Gabriel Meseguer-Brocal*[1]

[1]Deezer Research, Paris, France [2]Nantes Université, Centrale Nantes, CNRS, LS2N, UMR 6004, F-44000 Nantes, France

***Abstract*—Contrastive learning and equivariant learning are effective methods for self-supervised learning (SSL) for audio content analysis. Yet, their application to music information retrieval (MIR) faces a dilemma: the former is more effective on tagging (e.g., instrument recognition) but less effective on structured prediction (e.g., tonality estimation); The latter can match supervised methods on the specific task it is designed for, but it does not generalize well to other tasks. In this article, we adopt a best-of-both-worlds approach by training a deep neural network on both kinds of pretext tasks at once. The proposed new architecture is a Vision Transformer with 1-D spectrogram patches (ViT-1D), equipped with two class tokens, which are specialized to different self-supervised pretext tasks but optimized through the same model: hence the qualification of self-supervised multi-class-token multitask (MT^2). The former class token optimizes cross-power spectral density (CPSD) for equivariant learning over the circle of fifths, while the latter optimizes normalized temperature-scaled cross-entropy (NT-Xent) for contrastive learning. MT^2 combines the strengths of both pretext tasks and outperforms consistently both single-class-token ViT-1D models trained with either contrastive or equivariant learning. Averaging the two class tokens further improves performance on several tasks, highlighting the complementary nature of the representations learned by each class token. Furthermore, using the same single-linear-layer probing method on the features of last layer, MT^2 outperforms MERT on all tasks except for beat tracking; achieving this with 18x fewer parameters thanks to its multitasking capabilities. Our SSL benchmark demonstrates the versatility of our multi-class-token multitask learning approach for MIR applications.**

1. INTRODUCTION

Self-supervised learning (SSL) consists in designing a *pretext task* to train deep neural networks from unlabeled data. The promise of SSL is that, although this task does not reflect a real-world use case, its resolution is transferable to *downstream tasks* with little or no supervised fine-tuning [1]. SSL has recently found many applications in the context of music information retrieval (MIR) [2].

Contrastive learning is arguably the simplest kind of SSL, with CLMR [3] and MULE [4] being two examples for MIR. Given an audio segment, named *anchor*, these methods extract a *positive sample*, i.e., a neighboring segment from the same music track or an artificially modified version of the anchor. Segments from a distinct music track from the anchor are regarded as *negative samples*. Recent improvements to contrastive learning have focused on improving the random sampling of negatives in terms of effectiveness and interpretability [5], [6]. Yet, because it assigns the same importance to all negatives, contrastive learning is inefficient for structured downstream tasks such as tonality estimation [7].

Equivariant learning consists in regressing the parameter underlying a certain artificial transformation by learning to correlate two transformed copies of the same unlabeled audio segment. Crucially, these artificial transformations are tailored to the downstream task: e.g., time shifts for beat tracking [8], pitch shifts for fundamental frequency estimation [9], [10], time warps for tempo estimation [11], and so forth. Compared to contrastive learning, equivariant learning eliminates the need to carefully craft positive and negative samples for each anchor. Equivariant pretext tasks have the drawback of reducing the representation to a single degree of freedom.

Masked language modeling (MLM) involves predicting a masked word based on its context. Originally introduced as BERT in NLP, this pretext task has been adapted to music in models like MERT [12] and MusicFM [13]. While effective on downstream tasks, these models rely on complex SSL techniques, such as teacher-student distillation with exponential moving averages, which require costly, trial-and-error hyperparameter tuning. Furthermore, MLM requires large neural network architectures: e.g., 95M parameters for MERT.

Meanwhile, there is a pressing need for multitask MIR [14]–[21]. Yet, SSL for MIR faces a conundrum: equivariant learning achieves state-of-the-art (SOTA) performance on structured downstream tasks, contrastive learning excels in others, and MLM models reach SOTA performance across most tasks, though at a high computational cost.

In this article, we resolve this conundrum by training a neural network with only 5.3M parameters on multiple pretext tasks at once: i.e., a contrastive and an equivariant learning task. The key idea is to compute loss functions of each self-supervised pretext task over different *class tokens*, which are prepended to the *sequence tokens* of a Vision Transformer with 1-D patches (ViT-1D) [22]. Hence the proposed name: multi-class-token multitask (MT^2) ViT-1D. By averaging both class tokens for downstream tasks, MT^2 outperforms single-class-token monotask ViT-1D models with the same number of parameters, each optimized with either contrastive or equivariant loss, and the MLM model MERT's last-layer representation while having 18x fewer parameters. Interestingly, the advantage of MT^2 is not limited to class tokens but is also observable in sequence tokens, which outperform MERT in chord estimation. The code and model weights are publicly available at https://github.com/deezer/mt2.

2. METHODS

Our key contributions, multi-class-token multitasking (MT^2) and token-based downstream usage, are detailed in Sections 2.5 and 2.6, with an overview in Figure 1. ViT-Mel (Section 2.1) and ViT-CQT (Section 2.3), components of MT^2, also serve as baselines in Section 3.

2.1. Vision Transformer with mel-frequency patches (ViT-Mel)

ViT-Mel processes audio segments of four seconds, sampled at 16 kHz. We compute a mel-frequency spectrogram with 128 bins and a frame rate of 31.5 Hz. This representation is commonly used in MIR for contrastive learning in tagging tasks [2], [4]. Each spectrogram frame is a 1-D vector, indexed by mel-frequency. We normalize this vector and apply a learnable linear layer, producing a patch sequence $\boldsymbol{x}_{\mathrm{mel}} = [\boldsymbol{x}_1, ..., \boldsymbol{x}_T]$ in dimension $d_{\mathrm{e}} = 192$, with $T = 126$ being the sequence length.

We build a Vision Transformer with 1-D patches (ViT-1D) [22] with an embedding dimension of 192, a depth of 12 blocks, and three attention heads per block, which is the smallest ViT version proposed in the original paper [23]. We prepend a learnable *class token* $\boldsymbol{c}^{\mathrm{in}}_{\mathrm{cont}}$ to the patch sequence [24], defined as:

$$\boldsymbol{c}^{\mathrm{in}}_{\mathrm{cont}} = \tilde{\boldsymbol{c}}^{\mathrm{in}}_{\mathrm{cont}} + \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{x}_t^{mel}. \quad (1)$$

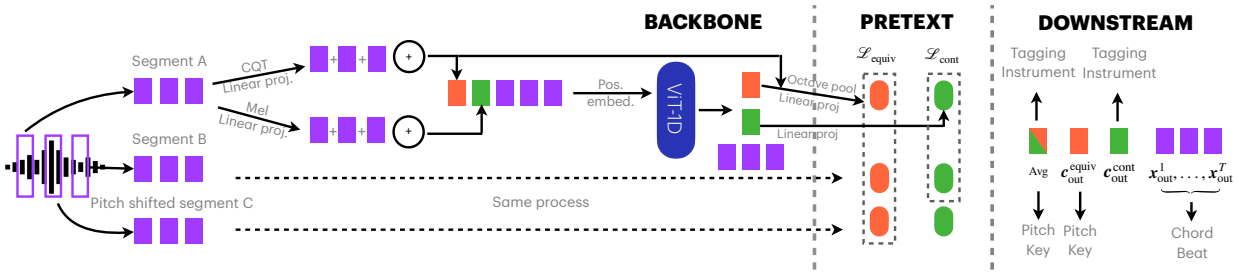

Fig. 1: Overview of the proposed method for mult-class-token multitask self-supervised learning with Vision Transformer with 1-D patches (MT^2).

where $\tilde{\boldsymbol{c}}^{\text{in}}_{\text{cont}}$ are learnable parameters, and "cont" refers to "contrastive", detailed in Section 2.2.

This yields the input sequence as $[\boldsymbol{c}^{\text{in}}_{\text{cont}}, \boldsymbol{x}^{\text{in}}_1, \ldots, \boldsymbol{x}^{\text{in}}_T]$. Likewise, the output sequence has the form $[\boldsymbol{c}^{\text{out}}_{\text{cont}}, \boldsymbol{x}^{\text{out}}_1, \ldots, \boldsymbol{x}^{\text{out}}_T]$. We apply a 2-D positional encoding over time and mel-frequency to the input sequence. In the ViT-1D, the *class token* is processed in the same way as sequence tokens, with shared weight in the MLP layers following the attention blocks. The *class token* serves as a summarization of the sequence, as proposed in [23]. Before applying the loss, a learnable linear layer maps $\boldsymbol{c}^{\text{out}}_{\text{cont}}$ onto a vector $\boldsymbol{z}_{\text{cont}}$ of dimension 512.

2.2. Normalized temperature-scaled cross-entropy (NT-Xent)

We train ViT-Mel with self-supervised contrastive learning. We extract two disjoint segments A and B from the same song and regard them as an anchor and a positive sample. In the same batch, extracted segments from other songs are considered as negative samples for this anchor [4]. The contrastive loss function is normalized temperature-scaled cross-entropy (NT-Xent), defined as :

$$\mathcal{L}^{\text{cont}}_{A,B}(\boldsymbol{w}) = -\log \frac{\exp(\text{sim}(\boldsymbol{z}^{\text{cont}}_A, \boldsymbol{z}^{\text{cont}}_B)/\tau)}{\sum_{k\neq B}\exp(\text{sim}(\boldsymbol{z}^{\text{cont}}_A, \boldsymbol{z}^{\text{cont}}_k)/\tau)} \tag{2}$$

where sim denotes the cosine similarity, $\tau = 0.1$ is a temperature hyperparameter, and $\boldsymbol{w}$ are the trainable weights of ViT-Mel.

2.3. Vision Transformer with CQT patches (ViT-CQT)

Alternatively, we build a ViT on the constant-Q transform (CQT) with $Q = 12$ bins per octave over $J = 8$ octaves. CQT is commonly used as representations for harmonic related tasks [25]–[27]. We make the same design choices as for ViT-Mel regarding embedding dimension d_{e}, sequence length T, ViT-1D architecture (Section 2.1). The CQT input sequence is denoted as $\boldsymbol{x}_{\text{cqt}} = [\boldsymbol{x}_1, ..., \boldsymbol{x}_T]$. We prepend a *class token* in the same way as for ViT-Mel, initialized with learnable parameters and the average of $\boldsymbol{x}_{\text{cqt}}$.

At the output of ViT-1D, a residual connection is added from the average of $\boldsymbol{x}_{\text{cqt}}$ to $\boldsymbol{c}^{\text{out}}_{\text{equiv}}$("equiv" refers to "equivariant", detailed in Section 2.4), obtaining:

$$\tilde{\boldsymbol{c}}^{\text{out}}_{\text{equiv}} = \boldsymbol{c}^{\text{out}}_{\text{equiv}} + \frac{1}{T}\sum_{t=1}^{T}\boldsymbol{x}^{\text{cqt}}_t. \tag{3}$$

We attach a learnable linear layer to project $\tilde{\boldsymbol{c}}^{\text{out}}_{\text{equiv}}$ into dimension of $QJ = 84$, then we reduce the dimension to $Q = 12$ via octave pooling. This is similar to the extraction of chroma features in MIR, except that our averaging procedure operates on a learned representation as opposed to directly on the CQT. After averaging, we apply a softmax layer, yielding a nonnegative vector $\boldsymbol{z}_{\text{equiv}}$ in dimension 12 whose entries sum to one.

2.4. Cross-power spectral density (CPSD)

We train ViT-CQT with self-supervised equivariant learning. Cross-power spectral density (CPSD), introduced in STONE, is used to structure the embedding space in a circularly equivariant manner, resembling the circle of fifths [25]. Let $\hat{\boldsymbol{z}}$ be the discrete Fourier transform over $\mathbb{Z}_{12}$ of $\boldsymbol{z}_{\text{equiv}}$. The CPSD of $\boldsymbol{z}^{\text{equiv}}_A$ and $\boldsymbol{z}^{\text{equiv}}_B$ is given by $\hat{\boldsymbol{z}}_A\hat{\boldsymbol{z}}^*_B$, where the asterisk denotes the complex conjugate. Given a hyperparameter $\omega \in \mathbb{Z}$ and a pitch class interval parameter $k \in \mathbb{Z}$, we consider the following function:

$$\mathcal{D}_k(\boldsymbol{z}_A, \boldsymbol{z}_B) = \frac{1}{2}\left|e^{-2\pi \text{i}\omega k/12} - \hat{\boldsymbol{z}}_A[\omega]\hat{\boldsymbol{z}}^*_B[\omega]\right|^2. \tag{4}$$

By setting $\omega = 7$, $\mathcal{D}_k$ is a differentiable distance function over the circle of fifths. Indeed, by property of the softmax and since 7 is coprime with 12, $\mathcal{D}_k$ is zero if and only if $\boldsymbol{z}^{\text{equiv}}_A$ and $\boldsymbol{z}^{\text{equiv}}_B$ have a single nonzero entry and differ by a circular shift of k semitones. Assuming that a musical piece does not modulate between segments A and B, the minimization of $\mathcal{D}_0(\boldsymbol{z}_{\text{A}}, \boldsymbol{z}_{\text{B}})$ with respect to trainable weights $\boldsymbol{w}$ of ViT-CQT is an equivariant pretext task for self-supervised tonality estimation. Moreover, assuming that a third segment C is available and is known to be k semitones apart from A and B, then it is judicious to minimize $\mathcal{D}_k(\boldsymbol{z}_{\text{A}}, \boldsymbol{z}_{\text{C}})$ and $\mathcal{D}_k(\boldsymbol{z}_{\text{B}}, \boldsymbol{z}_{\text{C}})$. In practice, we obtain C by artificial frequency transposition of A and randomize the parameter k uniformly between -5 and +6 semitones. Thus, our loss function for self-supervised equivariant learning is:

$$\mathcal{L}^{\text{equiv}}_{A,B,C}(\boldsymbol{w}) = \mathcal{D}_0(\boldsymbol{z}_A, \boldsymbol{z}_B) + \mathcal{D}_k(\boldsymbol{z}_A, \boldsymbol{z}_C) + \mathcal{D}_k(\boldsymbol{z}_B, \boldsymbol{z}_C).$$

The loss function above has allowed a self-supervised convolutional network to match the supervised SOTA in the automatic classification of major and minor keys [26]. To our knowledge, our article is the first to apply CPSD-based equivariant learning to Transformers, which, unlike convolutional networks, are not inherently equivariant to transpositions. The skip connection before the final linear layer is crucial for faster convergence when learning the equivariant property.

2.5. Multi-class-token multitasking (MT^2)

We add patch sequences $\boldsymbol{x}_{\text{cqt}}$ (Section 2.3) and $\boldsymbol{x}_{\text{mel}}$ (Section 2.1) to produce a patch sequence $\boldsymbol{x}_{\text{p}} = \boldsymbol{x}_{\text{cqt}} + \boldsymbol{x}_{\text{mel}}$, to which we prepend two learnable class tokens: an equivariant class token $\boldsymbol{c}^{\text{in}}_{\text{equiv}} \in \mathbb{R}^{d_e}$ initialized with learnable parameters and the average of $\boldsymbol{x}_{\text{cqt}}$, and a contrastive class token $\boldsymbol{c}^{\text{in}}_{\text{cont}} \in \mathbb{R}^{d_e}$ initialized with learnable parameters and the average of $\boldsymbol{x}_{\text{mel}}$. Thus, the input token sequence to the transformer consists of the concatenation of the two class tokens followed by the frame-wise patch sequence $\boldsymbol{x}_{\text{p}}$. We use the same ViT-1D architecture as ViT-Mel and ViT-CQT: therefore, no additional parameters are introduced when training MT^2. Both class tokens are updated through the same attention layers and MLP layers, with no constraints beyond their respective losses, allowing each to selectively aggregate information.

The output token sequence of the last layer is denoted as $[\boldsymbol{c}^{\text{equiv}}_{\text{out}}, \boldsymbol{c}^{\text{cont}}_{\text{out}}, \boldsymbol{x}^1_{\text{out}}, ..., \boldsymbol{x}^T_{\text{out}}]$ where $\boldsymbol{c}^{\text{equiv}}_{\text{out}}$ and $\boldsymbol{c}^{\text{cont}}_{\text{out}}$ correspond to the output equivariant and contrastive tokens respectively. Before applying the losses, we have two distinct heads for the two class

tokens, same as described in Section 2.1 and 2.3. $\boldsymbol{c}_{\text{out}}^{\text{equiv}}$ added with its residual connection from $\boldsymbol{x}_{\text{cqt}}$ is passed through an octave pooling layer and softmax layer obtaining an output of 12 dimensions $\boldsymbol{z}^{\text{equiv}}$, while $\boldsymbol{c}_{\text{out}}^{\text{cont}}$ is linearly projected into a space of 512 dimensions $\boldsymbol{z}^{\text{cont}}$.

We optimize $\boldsymbol{z}^{\text{equiv}}$ using an equivariant loss $\mathcal{L}^{\text{equiv}}$, and $\boldsymbol{z}^{\text{cont}}$ using a contrastive loss $\mathcal{L}^{\text{cont}}$ applied across the batch. Recent publications in computer vision have proposed similar transformer architectures in which multiple learnable tokens are prepended to the beginning of the sequence. [28] introduced register tokens, which are discarded during downstream training but serve to enhance the emergent properties observed in attention maps. MCTransformer+ [29] uses multiple class tokens with distinct supervision signals corresponding to different classes. These approaches have been shown to improve both the emergent properties of transformer representations and performance on downstream tasks. However, to the best of our knowledge, similar approaches have not yet been explored in a fully self-supervised manner, with different SSL losses, particularly in MIR. Unlike above work, both class tokens in MT^2 are optimized in a fully self-supervised manner, structured differently only through their respective losses, and both are required for downstream tasks.

2.6. Usage of tokens for downstream tasks

The output sequence of MT^2 consists of two class tokens, $\boldsymbol{c}_{\text{out}}^{\text{equiv}}$ and $\boldsymbol{c}_{\text{out}}^{\text{cont}}$, along with sequence tokens, $[\boldsymbol{x}_{\text{out}}^1, ..., \boldsymbol{x}_{\text{out}}^T]$. For downstream task training, we average both class tokens, as they are not explicitly trained to be disentangled and may capture complementary information. For comparison, we also train on specific tokens for corresponding tasks, as described in Section 3. Moreover, the sequence tokens exhibit emergent properties, as noted in computer vision [24] and music information retrieval [22]. Although these tokens are not directly optimized by the loss functions, we hypothesize that they capture frame-level information that both class tokens rely on to generate representations that can be optimized by losses, thus useful for downstream tasks.

3. APPLICATION TO MUSIC INFORMATION RETRIEVAL

3.1. Training details

We curate a subset of 100k songs from the catalog of a commercial music streaming service. We use a batch size of 128 pairs of 4-second segments, a base learning rate of 1×10^{-4} with a cosine decay until 5×10^{-7}, and train for 600 epochs with 512 steps per epoch. For all downstream tasks, we train only a single linear layer on top of the 192-dimensional class or sequence tokens, keeping the ViT-1D backbone frozen.

3.2. Global tasks

Global tasks are time-invariant: they require a single prediction per audio excerpt. We consider four well-known global tasks in MIR: multilabel music tagging, instrument recognition on isolated notes, key estimation, and pitch estimation on isolated notes.

Music tagging. We train, validate, and test on MagnaTagATune [30] with the same split as [31]. MagnaTagATune contains 25k songs from 230 artists, with multilabel annotation from 50 English tags. We probe on the contrastive class token $\boldsymbol{c}_{\text{out}}^{\text{cont}}$ and the average of both class tokens for MT^2, and the class token of both single-token monotask models (ViT-Mel and ViT-CQT). We evaluate the area under the receiver operating characteristic curve (ROC-AUC) and mean average precision (mAP) in their macro-aggregated versions.

Instrument recognition. We train, validate, and test on TinySOL [32] with a random 8:1:1 split. TinySOL contains 3k isolated notes from 14 instruments. We use the same probing method as for music tagging. We evaluate top-1 accuracy over 14 classes.

Table 1: Benchmark of four global tasks: multilabel music tagging, instrument recognition on isolated notes, key estimation, and pitch estimation on isolated notes. The top two rows are results of probing the class token of monotask single-token models: ViT-Mel (Section 2.1) with contrastive learning (Section 2.2) and ViT-CQT (Section 2.3) with equivariant learning (Section 2.4). The next three rows are our multi-token multitask model (MT^2, Section 2.5), after probing the contrastive token, the equivariant token, or an averaging of the two tokens (Section 2.6). The last row is probing of MERT's last layer representations, a pretrained masked language model (MLM) for music. See Section 3.2 for details about tasks and metrics.

MODEL	TOKEN	TAGGING		INSTRUMENT	KEY	PITCH
		MAP	ROC	ACC	MIREX	ACC
ViT-Mel	Cont.	0.334	0.852	0.904	0.577	0.973
ViT-CQT	Equiv.	0.229	0.779	0.548	0.733	0.973
MT^2	Cont.	0.390	0.884	0.925	–	–
MT^2	Equiv.	–	–	–	0.700	0.983
MT^2	Avg.	0.388	0.884	0.918	0.713	0.990
MERT	-	0.345	0.852	0.493	0.527	0.979

Key estimation. We train and validate on FMAKv2 [25] with a random 9:1 split. FMAKv2, an improved version of FMAK [33], contains 5k songs from the Free Music Archive [34]. We test on GiantSteps [35], which contains 604 electronic dance music tracks. We probe on the equivariant class token $\boldsymbol{c}_{\text{out}}^{\text{equiv}}$ and the average of both tokens for MT^2, and the class token of the both single-token monotask models. We evaluate the MIREX score, a weighted accuracy which assigns a lower penalty to predictions which are harmonically related to the ground truth [36].

Pitch estimation. We use TinySOL as for instrument recognition [32]. We use the same probing method as in key estimation. We evaluate top-1 classification accuracy over 82 semitone intervals.

Table 1 summarizes our results for all four global tasks.

3.3. Local tasks

Local tasks are time-equivariant: they require a prediction per frame with a frame rate typically higher than 1 Hz. We consider two well-established local tasks in MIR: beat tracking and chord estimation. Here, instead of probing on the class tokens, we probe on the sequence tokens $[\boldsymbol{x}_{\text{out}}^1, ..., \boldsymbol{x}_{\text{out}}^T]$ for ViT-Mel, ViT-CQT and MT^2. These sequence tokens are not directly optimized using any pretext task losses; however, they exhibit emergent properties that can be used for local tasks.

Beat tracking. We train and validate on the Ballroom dataset [37] with a 9:1 split. This dataset contains 698 songs. We test on GTZAN Rhythm [38], which contains 998 songs. Since beat tracking typically requires a higher frame rate than 31.5 Hz, we attach two independent heads to each $\boldsymbol{x}_{\text{out}}$, doubling the frame rate to 63 Hz. Additionally, we apply a standard smoothing method for beat tracking, where we increase the values of the two neighboring frames to 0.5 instead of 0. We evaluate the F_1-score with a tolerance window of ± 70 ms.

Chord estimation. We collect 124 songs from the Real World Computing Pop (RWC-POP) and Schubert Winterreise Dataset (SWD) [39], [40], limited to one performance per song. We train, validate, and test on this corpus with a random 8:1:1 split. We consider 24 classes of major and minor chords; exclude those that cannot be reduced to these classes (e.g., suspended chords); and include a "no chord" class, resulting in 25 classes. We evaluate frame-level classification accuracy at a rate of 31.5 Hz.

Table 2 summarizes our results for both local tasks.

Table 2: Benchmark on two local tasks: beat tracking and chord estimation, based on probing sequence tokens. See Table 1 for description of models. See Section 3.3 for details about tasks and metrics.

MODEL	TOKEN	BEAT F_1	CHORD ACC
ViT-Mel	Seq.	0.656	0.323
ViT-CQT	Seq.	0.515	0.542
MT^2	Seq.	0.698	0.447
MERT	-	0.786	0.392

4. DISCUSSION

4.1. Comparison with single-token monotask models

We compare MT^2 with two single-token monotask baseline models described in Section 2: ViT-Mel and ViT-CQT. ViT-Mel is a ViT-1D whose class token is initialized with a Mel-frequency spectrogram and trained using contrastive loss. ViT-CQT, by contrast, uses a CQT initialization and is optimized with the cross-power spectral density (CPSD) loss, an equivariant objective.

Compared to ViT-Mel, the equivariant class token of MT^2 shows large improvements on harmonically structured global tasks such as key and pitch estimation. For tagging and instrument recognition, which are better suited to contrastive pretraining, the contrastive class token of MT^2 still outperforms ViT-Mel, despite both being optimized with the same loss. This suggests that the multitask, multi-token framework enables beneficial inductive sharing between the two learning objectives, allowing the contrastive token to leverage information learned by the equivariant token.

ViT-CQT, being tailored for equivariant tasks through CPSD optimization, achieves strong results on harmonic tasks. However, its performance on other tasks is substantially lower, reflecting a known limitation of equivariant SSL: task-specific representations often do not generalize well. In contrast, while MT^2 shows a slight performance drop on key and chord estimation relative to ViT-CQT, it achieves considerably better results across all other tasks, demonstrating a favorable trade-off between specialization and generalization.

Moreover, we observe that averaging the two class tokens leads to 1) improved performance on key and pitch estimation, 2) matching performance on tagging, and 3) a slight drop on instrument recognition while still outperforming both single-token monotask models on this task. This suggests that the two class tokens encode complementary information and are not fully disentangled, highlighting the benefit of considering them in combination for certain downstream scenarios.

4.2. Comparison with MERT

MERT is a music representation model that achieves leading results across downstream tasks [12]. We use the public 95M-parameter backbone version (MERT-v1-95M). To ensure a fair evaluation against MT^2, we adopt the same datasets, splits, and probing described in Section 2. It is worth noting that MERT's original paper improves performance by using downstream-task-specific weight combinations of intermediate layers, alongside with dropout and a hidden layer of dimension 512. In contrast, we probe only a fixed layer output (the final layer) [3], [4], [13], apply no dropout, and use a single linear layer without hidden dimensions for both models. This might result in performance drop in downstream tasks for both models, however we focus on the benefits of multitasking by directly measuring the learned representation space for all downstream tasks.

Across all evaluated downstream tasks, MT^2 outperforms MERT under the same evaluation protocol, with the exception of beat tracking, while having 18× fewer parameters and a 4× smaller embedding space. In particular, it achieves gains in key and chord estimation, and instrument recognition. Interestingly, while MERT performs on par with ViT-Mel on non-harmonic tasks, MT^2 consistently outperforms on harmonic tasks, thanks to multitask learning via the equivariant class token. On beat tracking, MT^2 underperforms. This may be explained by the fact that the emergent properties in the sequence tokens are not explicitly guided by the loss functions.

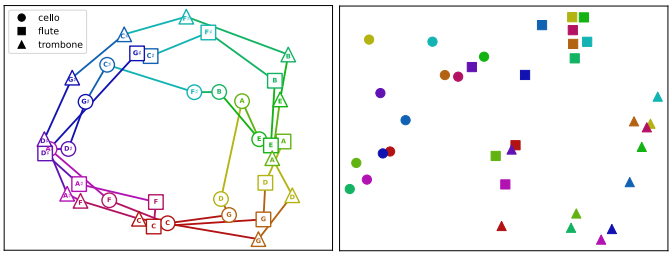

Fig. 2: PCA of MT^2 embeddings using the equivariant (left) and contrastive (right) class tokens. Shapes indicate instruments; colors indicate pitches. Left: pitches cluster across instruments and form a ring aligned with the circle of fifths (CoF). Right: samples cluster by instrument, showing pitch invariance. Edges (left) connect samples from the same instrument in the order of CoF, omitted on the right figure for clarity.

4.3. Principal component analysis

To interpret the roles of each class token in MT^2, we select three instruments playing 12 pitches each (from $\mathbb{A}_4$ to $\mathbb{G}\sharp_4$) from the TinySOL dataset, yielding 36 excerpts. These are processed by MT^2 after self-supervised training, without fine-tuning. We extract the 192-dimensional embeddings from both the contrastive and equivariant class tokens and reduce them to two dimensions using PCA. Figure 4.3 shows the resulting projections.

In the left subfigure, the equivariant token embeddings form a ring structure, where similar pitches from different instruments cluster together. The angular position aligns with the circle of fifths, and points for each instrument are connected in that order. Though not a perfect dodecagon, this structure generalizes across timbral variation. The subfigure on the right represents the contrastive class token. We observe that samples cluster primarily by instrument, regardless of pitch. This indicates that the contrastive token is relatively invariant to pitch and more sensitive to timbre.

These patterns suggest that assigning distinct self-supervised losses to each class token encourages the transformer to specialize: each class token captures complementary aspects of the sequence and structures the embedding space differently.

5. CONCLUSION

In the context of self-supervised learning (SSL) for music information retrieval (MIR), we have presented MT^2: a simple and scalable method to train the same ViT-1D model on multiple pretext tasks at once. The key idea is to allocate a different class token per task and to average them at the probing stage for downstream tasks. On four global tasks (music tagging, instrument recognition, key estimation, and pitch estimation), MT^2 matches or outperforms single-token competitors and MERT, a well-known system based on masked language modeling. While on local tasks, MT^2 trails behind MERT for the beat tracking and ViT-CQT for chord estimation, this may due to the sequence tokens being emergent but not directly optimized. We leave the improvement of these tasks to future work, potentially by further refining the equivariant objectives. Despite this limitation, our findings demonstrate the value of multitask SSL with multi-class-token transformer in MIR at the 100k-song scale.

REFERENCES

[1] R. Balestriero, M. Ibrahim, V. Sobal, A. Morcos, S. Shekhar, T. Goldstein, F. Bordes, A. Bardes, G. Mialon, Y. Tian *et al.*, "A cookbook of self-supervised learning," *arXiv preprint arXiv:2304.12210*, 2023.

[2] G. Meseguer-Brocal, D. Desblancs, and R. Hennequin, "An experimental comparison of multi-view self-supervised methods for music tagging," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024.

[3] J. Spijkervet and J. A. Burgoyne, "Contrastive learning of musical representations," in *Proc. International Society for Music Information Retrieval (ISMIR) Conference*, 2021.

[4] M. C. McCallum, F. Korzeniowski, S. Oramas, F. Gouyon, and A. F. Ehmann, "Supervised and unsupervised learning of audio representations for music understanding," in *Proc. International Society for Music Information Retrieval Conference (ISMIR)*, 2022.

[5] J. Guinot, E. Quinton, and G. Fazekas, "Leave-one-equivariant: Alleviating invariance-related information loss in contrastive music representations," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025, pp. 1–5.

[6] J. Choi, S. Jang, H. Cho *et al.*, "Towards proper contrastive self-supervised learning strategies for music audio representation," in *Proc. IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2022, pp. 1–6.

[7] J. S. Downie, A. F. Ehmann, M. Bay, and M. C. Jones, *The music information retrieval evaluation exchange: Some observations and insights*. Springer, 2010, pp. 93–115.

[8] D. Desblancs, V. Lostanlen, and R. Hennequin, "Zero-note samba: Self-supervised beat tracking," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023.

[9] B. Gfeller, C. Frank, D. Roblek, M. Sharifi, M. Tagliasacchi, and M. Velimirović, "SPICE: Self-supervised pitch estimation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1118–1128, 2020.

[10] A. Riou, S. Lattner, G. Hadjeres, and G. Peeters, "PESTO: Pitch estimation with self-supervised transposition-equivariant objective," in *Proc. International Society for Music Information Retrieval Conference (ISMIR)*, 2023.

[11] E. Quinton, "Equivariant self-supervision for musical tempo estimation," in *Proc. International Society for Music Information Retrieval Conference (ISMIR)*, 2022.

[12] Y. Li, R. Yuan, G. Zhang, Y. Ma, X. Chen, H. Yin, C. Lin, A. Ragni, E. Benetos, N. Gyenge, R. Dannenberg, R. Liu, W. Chen, G. Xia, Y. Shi, W. Huang, Y. Guo, and J. Fu, "Mert: Acoustic music understanding model with large-scale self-supervised training," in *Proc. International Conference on Learning Representations (ICLR)*, 2023.

[13] M. Won, Y.-N. Hung, and D. Le, "A foundation model for music informatics," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 1226–1230.

[14] J. Weston, S. Bengio, and P. Hamel, "Multi-tasking with joint semantic spaces for large-scale music annotation and retrieval," *Journal of New Music Research*, vol. 40, no. 4, pp. 337–348, 2011.

[15] P. Hamel, M. E. P. Davies, K. Yoshii, and M. Goto, "Transfer learning in MIR: Sharing learned latent representations for music audio classification and similarity," in *Proc. International Society for Music Information Retrieval Conference (ISMIR)*, 2013.

[16] X. Hu, J. H. Lee, D. Bainbridge, K. Choi, P. Organisciak, and J. S. Downie, "The MIREX grand challenge: A framework of holistic user-experience evaluation in music information retrieval," in *Proc. International Society for Music Information Retrieval Conference (ISMIR)*, 2014.

[17] R. E. P. Scholz, G. L. Ramalho, and G. Cabral, "Cross task study on MIREX recent results: An index for evolution measurement and some stagnation hypotheses," in *Proc. International Society for Music Information Retrieval Conference (ISMIR)*, 2016.

[18] Y.-N. Hung, Y.-A. Chen, and Y.-H. Yang, "Multitask learning for frame-level instrument recognition," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 381–385.

[19] J. Kim, J. Urbano, C. C. S. Liem, and A. Hanjalic, "One deep music representation to rule them all? a comparative analysis of different representation learning strategies," *Neural Computing and Applications*, vol. 32, no. 4, pp. 1067–1093, 2020.

[20] Y. Singh and A. Biswas, "Multitask learning based deep learning model for music artist and language recognition," in *Proc. Workshop on Speech and Music Processing (SMP)*, 2021.

[21] J. Turian, J. Shier, H. R. Khan, B. Raj, B. W. Schuller, C. J. Steinmetz, C. Malloy, G. Tzanetakis, G. Velarde, K. McNally *et al.*, "HEAR: Holistic evaluation of audio representations," in *Proceeedings of the Neural Information Processing Systems Conference (NeurIPS), Competitions and Demonstrations Track*. PMLR, 2022, pp. 125–145.

[22] Y. Kong, G. Meseguer-Brocal, V. Lostanlen, M. Lagrange, and R. Hennequin, "Emergent musical properties of a transformer under contrastive self-supervised learning," *International Society for Music Information Retrieval Conference (ISMIR 2025)*, 2025.

[23] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. International Conference on Learning Representations (ICLR)*, 2021.

[24] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," in *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 9650–9660.

[25] Y. Kong, V. Lostanlen, G. Meseguer-Brocal, S. Wong, M. Lagrange, and R. Hennequin, "STONE: Self-supervised tonality estimator," *Proc. International Society for Music Information Retrieval Conference (ISMIR)*, 2024.

[26] Y. Kong, G. Meseguer-Brocal, V. Lostanlen, M. Lagrange, and R. Hennequin, "S-key: Self-supervised learning of major and minor keys from audio," in *ICASSP 2025 - IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025, pp. 1–5.

[27] R. Bittner, B. McFee, J. Salamon, P. Li, and J. Bello, "Deep salience representations for f_0 estimation in polyphonic music," in *Proc. 18th International Society for Music Information Retrieval Conference (ISMIR)*, Suzhou, China, Oct. 2017.

[28] T. Darcet, M. Oquab, J. Mairal, and P. Bojanowski, "Vision transformers need registers," in *Proc. International Conference on Learning Representations (ICLR)*, 2024.

[29] L. Xu, M. Bennamoun, F. Boussaid, H. Laga, W. Ouyang, and D. Xu, "Mctformer+: Multi-class token transformer for weakly supervised semantic segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 12, pp. 8380–8395, 2024.

[30] E. Law, K. West, M. I. Mandel, M. Bay, and J. S. Downie, "Evaluation of algorithms using games: The case of music tagging," in *Proc. International Society for Music Information Retrieval Conference (ISMIR)*, 2009, pp. 213–218.

[31] J. Lee, J. Park, K. L. Kim, and J. Nam, "Sample-level deep convolutional neural networks for music auto-tagging using raw waveforms," in *Proc. International Conference on Sound and Music Computing (SMC)*, 2017.

[32] C. E. Cella, D. Ghisi, V. Lostanlen, F. Lévy, J. Fineberg, and Y. Maresz, "Orchideasol: A dataset of extended instrumental techniques for computer-aided orchestration," in *Proc. International Computer Music Conference (ICMC)*, 2020.

[33] S. Wong and G. Hernandez, "Fmak: A dataset of key and mode annotations for the free music archive — extended abstract," in *Proc. 24th International Society for Music Information Retrieval (ISMIR) Late-Breaking/Demo Papers*, Milan, Italy, 2023.

[34] M. Defferrard, K. Benzi, P. Vandergheynst, and X. Bresson, "Fma: A dataset for music analysis," *Proc. International Society for Music Information Retrieval Conference (ISMIR)*, 2017.

[35] P. Knees, Ángel Faraldo, P. Herrera, R. Vogl, S. Böck, F. Hörschläger, and M. L. Goff, "Two datasets for tempo estimation and key detection in electronic dance music annotated from user corrections," in *Proc. International Society for Music Information Retrieval Conference (ISMIR)*, 2015.

[36] C. Raffel, B. McFee, E. J. Humphrey, J. Salamon, O. Nieto, D. Liang, D. P. Ellis, and C. C. Raffel, "mir_eval: A transparent implementation of common mir metrics." in *Proc. International Society for Music Information Retrieval Conference (ISMIR)*, vol. 10, 2014, p. 2014.

[37] F. Gouyon and S. Dixon, "A review of rhythm description systems," in *Proc. International Society for Music Information Retrieval Conference (ISMIR)*, 2004.

[38] U. Marchand, Q. Fresnel, and G. Peeters, "Gtzan-rhythm: Extending the gtzan test-set with beat, downbeat and swing annotations," *Proc. International Society for Music Information Retrieval Late-breaking/Demo (ISMIR-LBD)*, 2015.

[39] C. Weiß, F. Zalkow, V. Arifi-Müller, M. Müller, H. V. Koops, A. Volk, and H. G. Grohganz, "Schubert winterreise dataset: A multimodal scenario for

music analysis," *Journal on Computing and Cultural Heritage (JOCCH)*, vol. 14, no. 2, pp. 1–18, 2021.

[40] M. Goto and H. Hashiguchi, "RWC music database: Popular, classical, and jazz music databases," *Proc. International Society for Music Information Retrieval Conference (ISMIR)*, 2002.