SKY: Self-supervised Learning of Major and Minor Keys from Audio

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

SKY: Self-supervised Learning of Major and Minor Keys from Audio

Yuexuan Kong[1,2], Gabriel Meseguer-Brocal[1], Vincent Lostanlen[2], Mathieu Lagrange[2], Romain Hennequin[1]
[1]*Deezer Research*
Paris, France
[2]*Nantes Université, Centrale Nantes, CNRS, LS2N, UMR 6004*
F-44000 Nantes, France

***Abstract*—STONE, the current method in self-supervised learning for tonality estimation in music signals, cannot distinguish relative keys, such as C major versus A minor. In this article, we extend the neural network architecture and learning objective of STONE to perform self-supervised learning of major and minor keys (SKY). Our main contribution is an auxiliary pretext task to STONE, formulated using transposition-invariant chroma features as a source of pseudo-labels. SKY matches the supervised state of the art in tonality estimation on FMAKv2 and GTZAN datasets while requiring no human annotation and having the same parameter budget as STONE. We build upon this result and expand the training set of SKY to a million songs, thus showing the potential of large-scale self-supervised learning in music information retrieval.**

***Index Terms*—music key estimation, self-supervised learning, music information retrieval**

I. Introduction

Variations in tonality tend to elicit sensations of surprise among music listeners [1]. Characterizing these variations is a long-standing topic in music information retrieval (MIR), with MIREX serving as a standard evaluation framework in the case of Western tonal harmony [2]. Yet, despite the interest in deep convolutional networks (convnets) in MIR [3], they depend on a collection of expert annotations for supervised learning. This is at odds with so-called *implicit learning* in humans: explicit understanding of erudite concepts of music theory is not necessary to perceive harmonic contrast. Hence, we question the need for supervision in machine learning for tonality estimation.

An alternative paradigm, known as self-supervised learning (SSL), has found promising applications into MIR [4]. The gist of SSL is to formulate a *pretext task*; i.e., one in which the correct answer may be inexpensively obtained from audio data. While some SSL systems have general-purpose pretext tasks and require supervised fine-tuning [5]–[8], others are tailored for specific downstream tasks: e.g., the estimation of pitch [9], [10], tempo [11], [12], beat [13], drumming patterns [14], and structure [15].

Very recently, a pretext task has been proposed for tonality estimation, as part of two SSL models:STONE, a key signature estimator, and its variant 24-STONE, the only existing self supervised key signature and mode estimator [16]. However, STONE is incomplete in the sense that it is sensitive to modulations within a given key signature: for example, STONE may distinguish C major from A major or from C minor, but not from A minor. On the other hand, 24-STONE, as a first proposition toward self-supervised key signature and mode estimator, underperforms by 2% to a supervised baseline, and by 15% when compared to models incorporating supervision. The issue of coming up with an SSL technique which could classify key signatures as well as major and minor modes that can achieve comparable performance as supervised models remains as an open problem.

In this article, we present SKY, the first SSL model which learns to represent both the distinction between key signatures and modes. Given that major and minor modes are the two most representative modes in western music, in this paper, we limit mode classification only to major and minor modes, which is often the case in literature [17]. The main idea behind SKY is to form pseudo-labels for the mode classification by comparing the chroma features which correspond to the root notes of the relative major and minor scales. To identify these root notes, we rely on self-supervised knowledge about key signatures, as obtained via a STONE-like pretext task. The originality of SKY is to re-inject this knowledge into the formulation of a finer-grained task. For simplicity and efficiency, our convnet optimizes both tasks at once, via a structured output for 24-class classification: 12 key signatures and two modes.

Our main finding is that SKY achieves a MIREX score [2] of 72.0% on the FMAKv2 dataset, outperforming the self-supervised state of the art (SOTA) of 57.9% held by 24-STONE with the same number of parameters and training samples (60k songs). Scaling up SSL to 1M songs brings the MIREX score of SKY up to 73.2%, on par with the *supervised* SOTA (73.1%) of [17]. We expand our MIREX-compliant benchmark to three other datasets: GTZAN, GiantSteps, and Schubert Winterreise Dataset (SWD). Although key classification remains challenging for certain genres (*e.g.*, blues, jazz, and hip-hop), SKY is the first SSL method which matches or outperforms supervised deep learning for this task with no need for supervision.

II. Methods

Our proposed method builds on previous publication [16] whose key components are briefly presented in II-A and II-B. From II-C to II-F, we introduce novel contributions of SKY which replace the necessity of supervision in 24-STONE by self-supervision.

A. Structured prediction with ChromaNet

For each song in the training set, we extract two disjoint time segments, denoted by A and B. We compute their constant-Q transforms (CQT) with $Q = 12$ bins per octave and center frequencies ranging between 27.5 Hz and 8.37 kHz (99 bins). We denote the CQT of segment A by $\boldsymbol{x}_{\mathrm{A}}$ and idem for $\boldsymbol{x}_{\mathrm{B}}$. This is non-contrastive SSL because $\boldsymbol{x}_{\mathrm{A}}$ and $\boldsymbol{x}_{\mathrm{B}}$ are assumed to be in the same key.

To perform artificial pitch transposition, we crop CQT rows in $\boldsymbol{x}_{\mathrm{A}}$ to simulate a pitch transposition by c semitones for $0 \leq c \leq 15$: $T_c\boldsymbol{x}_{\mathrm{A}}[p,t] = \boldsymbol{x}_{\mathrm{A}}[p-c,t]$ for each $c \leq p < QJ$ where $J = 7$ octaves. Idem for $T_c\boldsymbol{x}_{\mathrm{B}}$. All CQTs after cropping result in $QJ = 84$ bins in total. As an example, $T_0\boldsymbol{x}_{\mathrm{A}}$ and $T_k\boldsymbol{x}_{\mathrm{A}}$ are assumed to have a pitch difference of k semitones.

We define a 2-D fully convnet $f_{\boldsymbol{\theta}}$ with trainable parameters $\boldsymbol{\theta}$, operating on $T_c\boldsymbol{x}_{\mathrm{A}}$ and $T_c\boldsymbol{x}_{\mathrm{B}}$ with $M = 2$ output channels and no pooling over the frequency dimension. Over each channel, we apply average pooling on the time dimension and batch normalization.

The matrix of learnable activations $\boldsymbol{f}_{\boldsymbol{\theta}}(T_c\boldsymbol{x}_{\mathrm{A}})$ has $QJ = 84$ rows and $M = 2$ columns. We sum this matrix across octaves, i.e., across

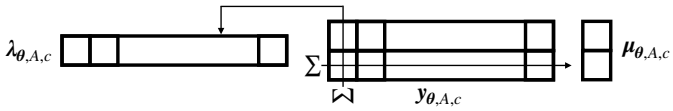

Fig. 1. Structured prediction: Summing $\boldsymbol{y}_{\boldsymbol{\theta},\mathrm{A,c}}$ over rows produces a pitch-equivariant component $\boldsymbol{\lambda}_{\boldsymbol{\theta},\mathrm{A,c}}$, summing $\boldsymbol{y}_{\boldsymbol{\theta},\mathrm{A,c}}$ per columns produces a pitch-invariant component $\boldsymbol{\mu}_{\boldsymbol{\theta},\mathrm{A,c}}$.

rows by Q semitones apart, and apply a softmax transformation over all $Q * M = 24$ entries.

This yields a matrix $\boldsymbol{y}_{\boldsymbol{\theta},\mathrm{A,c}}$ with $Q = 12$ rows and $M = 2$ columns whose entries are nonnegative and sum to one. We sum the columns of $\boldsymbol{y}_{\boldsymbol{\theta},\mathrm{A,c}}$, yielding $\boldsymbol{\lambda}_{\boldsymbol{\theta},\mathrm{A,c}}[q] = \sum_{m=0}^{M-1} \boldsymbol{y}_{\boldsymbol{\theta},\mathrm{A,c}}[q,m]$ a vector with Q nonnegative entries summing to one. Likewise over rows: $\boldsymbol{\mu}_{\boldsymbol{\theta},\mathrm{A,c}}[m] = \sum_{q=0}^{Q-1} \boldsymbol{y}_{\boldsymbol{\theta},\mathrm{A,c}}[q,m]$, a vector with M nonnegative entries summing to one. This is a kind of structured prediction: the learned representation $\boldsymbol{y}_{\boldsymbol{\theta},\mathrm{A,c}}$ has a pitch-equivariant component $\boldsymbol{\lambda}_{\boldsymbol{\theta},\mathrm{A,c}}$ and a pitch-invariant component $\boldsymbol{\mu}_{\boldsymbol{\theta},\mathrm{A,c}}$, as shown in Figure 1. Idem for $\boldsymbol{y}_{\boldsymbol{\theta},\mathrm{B,c}}$, $\boldsymbol{\lambda}_{\boldsymbol{\theta},\mathrm{B,c}}$, and $\boldsymbol{\mu}_{\boldsymbol{\theta},\mathrm{B,c}}$.

B. Cross-power spectral density (CPSD)

The cross-power spectral density (CPSD) of $\boldsymbol{\lambda}_{\boldsymbol{\theta},\mathrm{A,c}}$ and $\boldsymbol{\lambda}_{\boldsymbol{\theta},\mathrm{B,c}}$ is the product $\widehat{\boldsymbol{\lambda}}_{\boldsymbol{\theta},\mathrm{A,c}}[\omega]\widehat{\boldsymbol{\lambda}}^{*}_{\boldsymbol{\theta},\mathrm{B,c}}[\omega]$, where the hat denotes a discrete Fourier transform (DFT), the asterisk denotes a complex conjugation, and the discrete frequency variable ω is coprime with 12. We set $\omega = 7$ so that the phase of the CPSD coefficient denotes a key modulation over the circle of fifths (CoF)—see [16] for details.

Intuitively, while $\boldsymbol{\lambda}_{\boldsymbol{\theta},\mathrm{A,c}}$ is a one-hot encoding, $\widehat{\boldsymbol{\lambda}}_{\boldsymbol{\theta},\mathrm{A,c}}$ is a complex number of module 1 on the border of the CoF. Given an integer k, the CPSD of $\boldsymbol{\lambda}_{\boldsymbol{\theta},\mathrm{A,c}}$ and $\boldsymbol{\lambda}_{\boldsymbol{\theta},\mathrm{A,c+k}}$ is the difference of phases corresponding to a pitch modulation of k semitones on the CoF.

We define a CPSD-based function $\mathcal{D}_{\boldsymbol{\theta},c,k}$ which is equal to zero if and only if the vectors $\boldsymbol{\lambda}_{\boldsymbol{\theta},\mathrm{A,c}}$ and $\boldsymbol{\lambda}_{\boldsymbol{\theta},\mathrm{B,c+k}}$ contain a single nonzero coefficient and are equal up to circular shift by k:

$$\mathcal{D}_{\boldsymbol{\theta},c,k}(\boldsymbol{x}_{\mathrm{A}},\boldsymbol{x}_{\mathrm{B}}) = \frac{1}{2}\left|e^{-2\pi \mathrm{i}\omega k/Q} - \widehat{\boldsymbol{\lambda}}_{\boldsymbol{\theta},\mathrm{A,c}}[\omega]\widehat{\boldsymbol{\lambda}}^{*}_{\boldsymbol{\theta},\mathrm{B,c+k}}[\omega]\right|^2. \quad (1)$$

For any integer k and pair $\boldsymbol{x} = (\boldsymbol{x}_{\mathrm{A}},\boldsymbol{x}_{\mathrm{B}})$, $\mathcal{D}_{\boldsymbol{\theta},c,k}$ is differentiable with respect to ChromaNet weights $\boldsymbol{\theta}$. Hence, we define a CPSD-based loss function[1] which is parametrized by c and k:

$$\begin{aligned}\mathcal{L}_{\mathrm{CPSD}}(\boldsymbol{\theta}\,|\,\boldsymbol{x},c,k) = {}& \mathcal{D}_{\boldsymbol{\theta},c,0}(\boldsymbol{x}_{\mathrm{A}},\boldsymbol{x}_{\mathrm{B}}) \\ &+ \mathcal{D}_{\boldsymbol{\theta},c,k}(\boldsymbol{x}_{\mathrm{A}},\boldsymbol{x}_{\mathrm{A}}) \\ &+ \mathcal{D}_{\boldsymbol{\theta},c,k}(\boldsymbol{x}_{\mathrm{B}},\boldsymbol{x}_{\mathrm{A}}). \end{aligned} \quad (2)$$

In equation (2), the first term encourages the model $\boldsymbol{f}_{\boldsymbol{\theta}}$ to be invariant to the permutation of $\boldsymbol{x}_{\mathrm{A}}$ and $\boldsymbol{x}_{\mathrm{B}}$, while the second and third term encourage it to be equivariant to the pitch interval k. As [16] points out, all three terms are indispensable for an efficient optimization of the model without collapsing into a uniform or constant distribution.

C. Pseudo-labeling of mode

STONE has shown that training a ChromaNet to minimize $\mathcal{L}_{\mathrm{CPSD}}$ produces a pitch-equivariant representation which is a sparse nonnegative vector in dimension Q. We elaborate on this prior work to build a self-supervised approximate predictor of key signature, based on the pitch-equivariant component $\boldsymbol{\lambda}_{\boldsymbol{\theta}}$ for both segments A and B:

$$q_{\max}(\boldsymbol{\theta}\,|\,\boldsymbol{x}) = \arg\max_{0\leq q<Q} (\boldsymbol{\lambda}_{\boldsymbol{\theta},\mathrm{A,c}}[q] + \boldsymbol{\lambda}_{\boldsymbol{\theta},\mathrm{B,c}})[q]). \quad (3)$$

[1]In this paper, we use the vertical bar notation to clearly separate neural network parameters on the left versus data and random values on the right.

Our postulate is that, if $\mathcal{L}_{\mathrm{CPSD}}(\boldsymbol{\theta})$ is low and $\boldsymbol{x}$ is in a major key, $q_{\max}(\boldsymbol{\theta}\,|\,\boldsymbol{x})$ on the CQT scale corresponds to its root pitch class.

We compute a pitch class profile (PCP) for $\boldsymbol{x}$ by averaging its CQT across octaves, along time, and across segments A and B:

$$\boldsymbol{u}(\boldsymbol{x})[q] = \frac{1}{2}\sum_{j=0}^{J-1}\sum_{t=0}^{\tau-1}\big(\boldsymbol{x}_{\mathrm{A}}[Qj+q,t] + \boldsymbol{x}_{\mathrm{B}}[Qj+q,t]\big) \quad (4)$$

Without side information nor learning, $\boldsymbol{u}(\boldsymbol{x})$ would be a poor predictor of tonality, as it erases spectrotemporal dynamics in $\boldsymbol{x}$. However, when the key signature is known (e.g., no ♭ nor ♯), comparing the CQT energy of the root note of the major key (e.g., C) with that of the relative minor key (e.g., A) can achieve an accuracy of 79.4% in correctly determining the mode. Our main idea for this paper is to use the key signature predictor $q_{\max}(\boldsymbol{\theta})$ as side information to improve pretext task design based on $\boldsymbol{u}(\boldsymbol{x})$.

We look up the entry $u_{\mathrm{maj}}(\boldsymbol{\theta}\,|\,\boldsymbol{x},c) = \boldsymbol{u}(T_c\boldsymbol{x})[q_{\max}(\boldsymbol{\theta}\,|\,\boldsymbol{x})]$, where $T_c\boldsymbol{x}$ is a shorthand for $(T_c\boldsymbol{x}_{\mathrm{A}}, T_c\boldsymbol{x}_{\mathrm{B}})$. Its value may be interpreted as the acoustical energy at the root pitch class under the assumption that the song is in a major key. Conversely, we look up $u_{\min}(\boldsymbol{\theta}\,|\,\boldsymbol{x},c) = \boldsymbol{u}(T_c\boldsymbol{x})[(q_{\max}(\boldsymbol{\theta}\,|\,\boldsymbol{x}) - 3) \bmod Q]$, i.e., idem under the assumption that the song is in a minor key. Since $Q = 12$, the number 3 in the definition of $u_{\min}$ corresponds to a minor third, i.e., the interval between roots of relative keys. We define a pseudo-label $\boldsymbol{\nu}$ for SSL of mode according to a simple logical rule:

$$\boldsymbol{\nu}(\boldsymbol{\theta}\,|\,\boldsymbol{x},c) = \begin{cases}[1,0] & \text{if } (u_{\mathrm{maj}}(\boldsymbol{\theta}\,|\,\boldsymbol{x},c) > u_{\min}(\boldsymbol{\theta}\,|\,\boldsymbol{x},c)) \\ [0,1] & \text{otherwise.}\end{cases} \quad (5)$$

D. Binary cross-entropy (BCE) with pseudo-labels

Given $\boldsymbol{\nu}(\boldsymbol{\theta}\,|\,\boldsymbol{x},c)$ and k, we define a novel loss function:

$$\begin{aligned}\mathcal{L}_{\mathrm{SKY}}(\boldsymbol{\theta}\,|\,\boldsymbol{x},c,k) = {}& \mathrm{BCE}(\boldsymbol{\nu}(\boldsymbol{\theta}\,|\,\boldsymbol{x},c), \boldsymbol{\mu}_{\boldsymbol{\theta},\mathrm{A,c}}) \\ &+ \mathrm{BCE}(\boldsymbol{\nu}(\boldsymbol{\theta}\,|\,\boldsymbol{x},c), \boldsymbol{\mu}_{\boldsymbol{\theta},\mathrm{B,c}}) \\ &+ \mathrm{BCE}(\boldsymbol{\nu}(\boldsymbol{\theta}\,|\,\boldsymbol{x},c), \boldsymbol{\mu}_{\boldsymbol{\theta},\mathrm{A,c+k}}) \end{aligned} \quad (6)$$

where $\mathrm{BCE}(\boldsymbol{\nu},\boldsymbol{\mu}) = -\boldsymbol{\nu}[0]\log\boldsymbol{\mu}[0] - \boldsymbol{\nu}[1]\log\boldsymbol{\mu}[1]$ denotes binary cross-entropy. Intuitively, $\mathcal{L}_{\mathrm{SKY}}$ is low if and only if the structured predictions $\boldsymbol{f}_{\boldsymbol{\theta}}(T_c\boldsymbol{x}_{\mathrm{A}})$, $\boldsymbol{f}_{\boldsymbol{\theta}}(T_c\boldsymbol{x}_{\mathrm{B}})$, and $\boldsymbol{f}_{\boldsymbol{\theta}}(T_{c+k}\boldsymbol{x}_{\mathrm{A}})$ have large coefficients in the column corresponding to the pseudo-label $\boldsymbol{\nu}(\boldsymbol{\theta}\,|\,\boldsymbol{x},c)$.

Crucially, the equation above is different from the definition of $\mathcal{L}_{\mathrm{BCE}}$ in 24-STONE [16, Equation 16], which only involves pairwise BCE's between ChromaNet activations $\boldsymbol{\mu}_{\boldsymbol{\theta}}$.

While STONE is symmetric across columns, SKY breaks this asymmetry via the pseudo-labeling function $\boldsymbol{\nu}$, making it less susceptible to model collapse. This pseudo-labeling process replaced the indispensable supervision in 24-STONE.

E. Loss over batch-wise average of mode predictions

SSL training with $\mathcal{L}_{\mathrm{SKY}}$ faces a "cold start" problem in the sense that the pseudo-labeling function $\boldsymbol{\nu}$ is itself parametrized by the pitch equivariant component $\boldsymbol{\lambda}_{\boldsymbol{\theta}}$, therefore ChromaNet weights $\boldsymbol{\theta}$. During informal experiments, we have observed that penalizing $\boldsymbol{\theta}$ with $\mathcal{L}_{\mathrm{CPSD}}$ may not suffice to bootstrap the model from a random initial value. Against this issue, we assume that roughly half of the songs in each mini-batch of N songs $\mathbf{X} = (\boldsymbol{x}_n)_{n=0}^{N-1}$ are major, the other half being minor. We denote the corresponding batches of pitch transposition parameters by $\mathbf{C} = (\mathbf{C}[n])_n$ and $\mathbf{K} = (\mathbf{K}[n])_n$. We use $T_{\mathbf{C}}\mathbf{X}$ as a shorthand for $((T_{\mathbf{C}[n]}\mathbf{X}_{n,\mathrm{A}}, T_{\mathbf{C}[n]}\mathbf{X}_{n,\mathrm{B}}))_n$. We compute the batch-wise average of mode predictions as

$$\mu^{\mathrm{avg}}_{\boldsymbol{\theta}}(T_{\mathbf{C}}\mathbf{X}) = \frac{1}{N}\sum_{n=0}^{N-1}\sum_{L\in\{A,B\}}\boldsymbol{\mu}_{\boldsymbol{\theta}}(T_{\mathbf{C}[n]}\mathbf{X}_{n,\mathrm{L}})[0] \quad (7)$$

and derive the loss function: $\mathcal{L}_{\text{avg}}(\boldsymbol{\theta}|\ \boldsymbol{x}, \mathbf{C}) = (\mu_{\boldsymbol{\theta}}^{\text{avg}}(T_{\mathbf{C}}\mathbf{X}) - \frac{1}{2})^2$.

F. Self-supervised learning of major and minor keys (SKY)

Summing all three terms yields the training loss for SKY:

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta}|\ \mathbf{X}, \mathbf{C}, \mathbf{K}) = &\sum_{n=0}^{N-1} \mathcal{L}_{\text{CPSD}}(\boldsymbol{\theta}\,|\,\mathbf{X}_n, \mathbf{C}[n], \mathbf{K}[n]) \\ &+ \lambda_{\text{SKY}} \sum_{n=0}^{N-1} \mathcal{L}_{\text{SKY}}(\boldsymbol{\theta}\,|\,\mathbf{X}_n, \mathbf{C}[n], \mathbf{K}[n]) \\ &+ \lambda_{\text{avg}} \mathcal{L}_{\text{avg}}(\boldsymbol{\theta}\,|\,\mathbf{X}, \mathbf{C}). \end{aligned} \tag{8}$$

We set the hyperparameters λ_{BCE} and λ_{avg} so that all three terms in the loss $\mathcal{L}$ are of the same order of magnitude at the initialization: $\lambda_{\text{BCE}} = 1.5$ and $\lambda_{\text{avg}} = 15$.

III. Application

A. Training

STONE was trained on a corpus of 60k songs from the Deezer catalog. To offer a fair comparison, we begin by training SKY on the exact same dataset: see IV-A and Table I. Later on, we scale up SSL training to 1M songs from Deezer: see IV-B and Table II.

We set the duration of segments A and B to 15 seconds. We randomize c uniformly between 0 and 15 semitones, k uniformly between -12 to 12 semitones and $0 \leq k + c \leq 15$. We train SKY for 50 epochs and use a batch size of 128 on the 60k-song corpus versus 100 epochs and a batch size of 256 on the 1M-song corpus. We use the AdamW optimizer with a learning rate of 0.001 and a cosine learning rate schedule preceded by a linear warm-up.

B. Calibration on C major and A minor scales

The necessity of calibrating two channels separately arises because the model sometimes reaches a local minimum where a shift of fifths exists between the two channels (e.g., C major has the same index as E minor, and as note C in CQT). In this local minimum, $\mathcal{L}_{\text{CPSD}}$ remains low, given that the fifths of a key are considered to be the closest among all keys except for the correct one. $\boldsymbol{\nu}$ would serve as a slightly less accurate pseudo-label than when the model is in its global minimum, however remains a relevant pseudo-label, as demonstrated by empirical results.

We create two synthetic samples, one in C major and another in A minor to calibrate two channels separately. This calibration step is similar to STONE [16] except that it operates on a structured output with two modes.

C. Self-supervised and supervised competitors

We compare SKY against three self-supervised systems:

- **Krumhansl** [18]. A template matching algorithm for CQT features in which major and minor templates are derived from psychoacoustic judgments, with no machine learning.
- **24-STONE** [16]. The self-supervised SOTA. It relies on CPSD for equivariance to key signature and on BCE for invariance to mode, with no pseudo-labels.
- **ν-STONE**. A simple new method which is an ad hoc procedure using a pre-trained STONE model [16]'s prediction of key signature and the rule-based heuristic ν (Section II-C) for mode prediction which requires no further training.

In addition, we compare SKY against the supervised SOTA:

- **madmom** [17]. An all-convolutional neural network, trained on a varied corpus (electronic dance music, pop/rock, and classical music) and made available as part of the madmom open-source software library for MIR [19, v0.16.1].

D. Evaluation datasets and metrics

We evaluate all systems on the following four datasets, which are labeled according to a taxonomy of 24 major and minor keys:

- **FMAKv2** [16]. 5,489 songs from the Free Music Archive (FMA), spread across 17 genres. It contains key annotation for each song and genre annotations for nearly half of them.
- **GTZAN** [20]. 837 songs from 9 genres. Only songs with a unique key are annotated, therefore no classical music is included.
- **GiantSteps** [21]. 604 two-minute excerpts of electronic dance music (EDM) from commercial songs.
- **SWD** [22]. 48 classical music pieces composed by Schubert. We only use the first 30s given that key modulations are common in classical music.

The MIREX score, as implemented in mir_eval, is weighted according to the tonal proximity between reference and prediction [23]. Key signature estimation accuracy (KSEA) assigns a full point to the prediction if it matches the reference and a half point if the prediction is one perfect fifth above or below the reference, and zero otherwise [16]. Mode accuracy assigns a full point if reference and prediction share the same mode (major or minor) and zero otherwise.

IV. Results

A. Self-supervised learning from 60k songs

We train all SSL methods on the same 60k-song corpus (see Section III-A) and compare them against a template matching algorithm (Krumhansl [18]) and the supervised SOTA [17].

Table I summarizes our results on FMAKv2, the largest dataset to date for evaluating tonality estimation. SKY outperforms the SSL SOTA (24-STONE) as well as Krumhansl's template matching algorithm. Furthermore, on all three metrics, the performance of SKY is within one percentage point of the supervised SOTA. Thus, SKY offers the first proof of feasibility for the value of SSL in full-fledged tonality estimation, i.e., with a taxonomy of 24 keys.

	MIREX (%)	KSEA (%)	mode acc. (%)
Krumhansl [18]	53.4	60.1	64.9
24-STONE [16]	57.9	78.0	62.2
ν-STONE	67.8	79.1	74.1
SKY (60k)	**72.1**	**80.3**	**79.0**
madmom [17]	73.1	81.3	79.3

TABLE I

CLASSIFICATION OF MAJOR AND MINOR KEYS IN THE FMAKV2 DATASET ACCORDING TO THREE METRICS: MIREX SCORE, KEY SIGNATURE ESTIMATION ACCURACY (KSEA) AND MODE ACCURACY. KRUMHANSL'S METHOD INVOLVES NO TRAINING, WHILE 24-STONE, ν-STONE, AND SKY ARE SELF-SUPERVISED ON THE SAME DATASET OF 60K SONGS. WE INCLUDE THE RESULTS OF THE MADMOM LIBRARY AS SUPERVISED STATE-OF-THE-ART FOR REFERENCE.

Breaking down the MIREX score into finer-grained metrics, we observe that the gap in performance between 24-STONE and ν-STONE is primarily attributable to a higher mode accuracy (62.2% versus 74.1%) rather than to a higher key signature estimation accuracy (KSEA, 78.0% versus 79.1%). This observation confirms that the rule-based procedure $\boldsymbol{\nu}$ (see Section II-C) is more effective for distinguishing a major key from its relative minor than the BCE-based loss initially developed for 24-STONE.

Unlike ν-STONE, SKY is trained from scratch to minimize a joint SSL objective (Equation (8)) in which $\boldsymbol{\nu}$ plays the role of a pseudo-labeling function. We posit that this joint optimization creates a virtuous circle: a lower value of the loss improves the informativeness of pseudo-labels, thus making the pretext task less ambiguous, and

so forth. Hence, the data-driven component in SKY is able to refine and surpass the ad hoc procedure in ν-STONE.

From ν-STONE to SKY, there is not only an improvement in terms of mode accuracy (74.1% versus 79.0%), but also in terms of KSEA (79.1% versus 80.3%). This seems to be a benefit of weight sharing and structured prediction in SKY.

B. Scaling up to 1M songs

Inspired by recent works on large-scale SSL for MIR [8], [24], we retrain SKY on a corpus of 1M songs from the Deezer catalog. Then, we evaluate both versions of SKY on FMAKv2 as well as three other annotated datasets: see Section III-D. Table II summarizes our findings. After SSL on 1M songs, SKY performs on-par with the supervised SOTA across all datasets. Scaling up the training set of SKY appears beneficial for three datasets out of four.

Dataset	FMAKv2	GTZAN	GiantSteps	SWD
#songs	5,489	837	604	40
SKY (60k)	72.1	70.9	**71.7**	89.0
SKY (1M)	**73.2**	**73.8**	71.2	**90.4**
madmom [17]	73.1	67.9	71.0	87.7

TABLE II
MIREX SCORE (%) OF SKY AFTER SELF-SUPERVISED TRAINING ON 60K OR 1M SONGS. We compare with the madmom package as supervised state of the art. Note: for madmom, we report a score that is lower than the one reported in the original paper [17], i.e., 74.6%, which might due to the different implementations used in madmom and in original paper.

C. Error analysis across genres

Figure 2 compares SKY versus the supervised SOTA across multiple datasets and genres. Within GTZAN, both methods achieve a MIREX score above 90% on *country* and below 50% on *blues*. In other words, the gap in MIREX score across genres is much greater than the gap between the two methods over GTZAN as a whole. Arguably, the MIREX taxonomy of 24 keys is inadequate for blues [25], [26]—likewise, to some extent, for jazz and hip-hop. We leave this important question to future work.

Moreover, the performance for *jazz* shows a large difference between FMAKv2 and Giantsteps. This might be due to the differing genre taxonomies and varying definitions of keys used by annotators [27].

With this caveat in mind, we observe that SKY outperforms the supervised SOTA on genres with diverse musical features: e.g., metal, jazz, and reggae. This suggests that SSL with SKY learns invariant representations of tonality. The only large downgrade from madmom to SKY is *old-time/historic*, a small subcorpus of 16 songs in FMAKv2. The small amount of data could lead to a noisy MIREX score.

D. Visualization of SKY embeddings

We interpret SKY via principal component analysis (PCA) of intermediate features after uniform averaging over time and across ChromaNet channels. As shown in Figure 3, songs in FMAKv2 form a ring pattern which is well explained by the circular progression of fifths, both for major keys (left) and minor keys (right). Crucially, PCA on CQT features does not show such interpretable patterns.

The circularity of key signatures in SKY embeddings results from equivariance in our pretext task design. This observation is reminiscent of foundational work on self-organizing maps for music cognition [28] and more recent work on unsupervised learning of octave equivalence [29]. Meanwhile, the originality of our finding is that it was obtained by analyzing an unlabled corpus of 1M songs, as opposed to subjective ratings [28] or monophonic sounds [29].

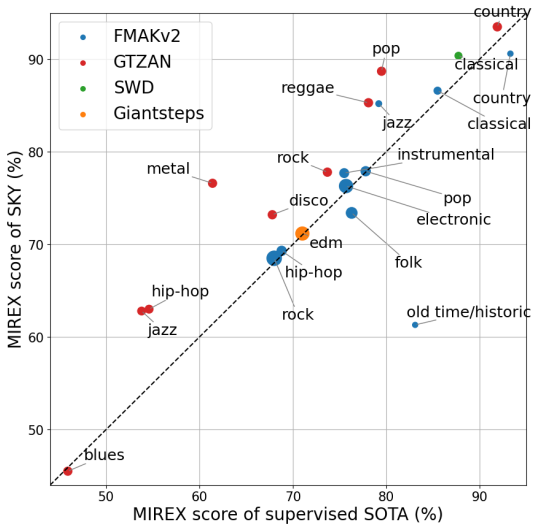

Fig. 2. Comparison between the supervised state of the art (x-axis) and SKY after self-supervised training on 1M songs (y-axis) in terms of MIREX score, across datasets and genres. The size of each marker is proportional to the number of songs in the corresponding subcorpus.

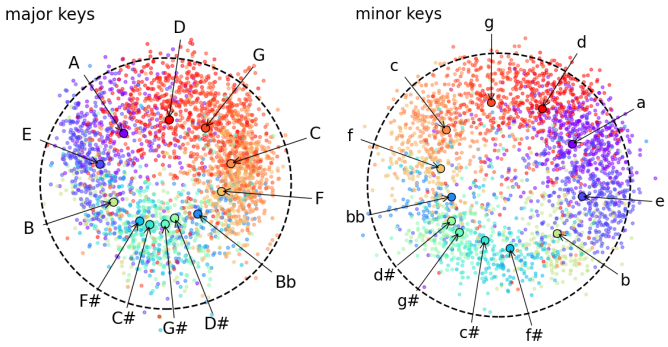

Fig. 3. 2-D visualization of FMAKv2 songs in major and minor keys after self-supervised embedding with SKY (trained on 1M songs) and principal component analysis (PCA). Hue indicates key on the circle of fifths, with key labels point at class centroids.

V. CONCLUSION

The promise of self-supervised learning (SSL) in music information retrieval is to harness large unlabeled music corpora to train deep neural networks with little or no annotation effort. In this article, we have presented SKY, an architecture and pretext task for self-supervised learning of 24 keys from audio. After SSL on 1M songs, SKY matches the supervised SOTA on four datasets. The main limitation behind SKY is that its structured prediction is limited to 24 major and minor keys, making it inadequate for certain genres. Still, the methodological contributions of SKY—namely, cross-power spectral density and pitch-invariant pseudo-labeling—could, in principle, apply to blues harmony and modal harmony, given appropriate training data and music-theoretical knowledge.

REFERENCES

[1] Richard Parncutt, *Psychoacoustic foundations of major-minor tonality*, MIT Press, 2024.

[2] J Stephen Downie, Andreas F Ehmann, Mert Bay, and M Cameron Jones, "The music information retrieval evaluation exchange: Some observations and insights," *Advances in music information retrieval*, pp. 93–115, 2010.

[3] Eric J. Humphrey, Juan P. Bello, and Yann Le Cun, "Feature learning and deep architectures: New directions for music informatics," *Journal of Intelligent Information Systems*, vol. 41, pp. 461–481, 2013.

[4] Shuo Liu, Adria Mallol-Ragolta, Emilia Parada-Cabaleiro, Kun Qian, Xin Jing, Alexander Kathan, Bin Hu, and Björn W. Schuller, "Audio self-supervised learning: A survey," *Patterns*, vol. 3, no. 12, 2022.

[5] Haohe Liu, Yi Yuan, Xubo Liu, Xinhao Mei, Qiuqiang Kong, Qiao Tian, Yuping Wang, Wenwu Wang, Yuxuan Wang, and Mark D. Plumbley, "AudioLDM 2: Learning holistic audio generation with self-supervised pretraining," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 2871–2883, 2024.

[6] Daisuke Niizumi, Daiki Takeuchi, Yasunori Ohishi, Noboru Harada, and Kunio Kashino, "BYOL for audio: Self-supervised learning for general-purpose audio representation," in *Proceedings of the IEEE International Joint Conference on Neural Networks (IJCNN)*, 2021.

[7] Janne Spijkervet and John Ashley Burgoyne, "Contrastive learning of musical representations," in *Proceedings of the International Society for Music Information Retrieval (ISMIR) Conference*, 2021.

[8] Yizhi Li, Ruibin Yuan, Ge Zhang, Yinghao Ma, Xingran Chen, Hanzhi Yin, Chenghao Xiao, Chenghua Lin, Anton Ragni, Emmanouil Benetos, Norbert Gyenge, Roger Dannenberg, Ruibo Liu, Wenhu Chen, Gus Xia, Yemin Shi, Wenhao Huang, Zili Wang, Yike Guo, and Jie Fu, "MERT: Acoustic music understanding model with large-scale self-supervised training," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024.

[9] Beat Gfeller, Christian Frank, Dominik Roblek, Matt Sharifi, Marco Tagliasacchi, and Mihajlo Velimirović, "SPICE: Self-supervised pitch estimation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1118–1128, 2020.

[10] Alain Riou, Stefan Lattner, Gaëtan Hadjeres, and Geoffroy Peeters, "PESTO: Pitch estimation with self-supervised transposition-equivariant objective," in *Proceedings from the International Society for Music Information Retrieval Conference (ISMIR)*, 2023.

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

[12] Antonin Gagneré, Slim Essid, and Geoffroy Peeters, "Adapting pitch-based self supervised learning models for tempo estimation," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 956–960.

[13] Dorian Desblancs, Vincent Lostanlen, and Romain Hennequin, "Zero-note samba: Self-supervised beat tracking," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023.

[14] Keunwoo Choi and Kyunghyun Cho, "Deep unsupervised drum transcription," in *Proceedings of the International Society for Music Information Retrieval (ISMIR) Conference*, 2019.

[15] Morgan Buisson, Brian Mcfee, Slim Essid, and Helene-Camille Crayencour, "Learning multi-level representations for hierarchical music structure analysis," in *Proceedings of the International Society for Music Information Retrieval (ISMIR)*, 2022.

[16] Yuexuan Kong, Vincent Lostanlen, Gabriel Meseguer-Brocal, Stella Wong, Mathieu Lagrange, and Romain Hennequin, "Stone: Self-supervised tonality estimator," *International Society for Music Information Retrieval Conference (ISMIR)*, 2024.

[17] Filip Korzeniowski and Gerhard Widmer, "Genre-agnostic key classification with convolutional neural networks," in *Proceedings of the International Society on Music Information Conference (ISMIR)*, 2018.

[18] Carol L. Krumhansl, *Cognitive foundations of musical pitch*, Oxford University Press, 2001.

[19] Sebastian Böck, Filip Korzeniowski, Jan Schlüter, Florian Krebs, and Gerhard Widmer, "Madmom: A new Python audio and music signal processing library," in *Proceedings of the 24th ACM international conference on Multimedia*, 2016, pp. 1174–1178.

[20] Cian Brien and Alexander Lerch, "Genre-specific key profiles," in *Proceedings of the International Computer Music Association Conference (ICMC)*, 2015.

[21] Ángel Faraldo Peter Knees and Richard Vogl, "Giantsteps key dataset," https://github.com/GiantSteps/giantsteps-key-dataset, 2015.

[22] Christof Weiß, Frank Zalkow, Vlora Arifi-Müller, Meinard Müller, Hendrik Vincent Koops, Anja Volk, and Harald 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.

[23] Colin Raffel, Brian McFee, Eric J Humphrey, Justin Salamon, Oriol Nieto, Dawen Liang, Daniel PW Ellis, and C Colin Raffel, "mir_eval: A Transparent Implementation of Common MIR Metrics.," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, 2014.

[24] Gabriel Meseguer-Brocal, Dorian Desblancs, and Romain Hennequin, "An experimental comparison of multi-view self-supervised methods for music tagging," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 1141–1145.

[25] Andrew Jaffe, *Something Borrowed Something Blue: Principles of Jazz Composition*, Advance Music, 2011.

[26] Ethan Hein, "Blues tonality," https://www.ethanhein.com/wp/2014/blues-tonality/, 2014.

[27] Bob L Sturm, "The gtzan dataset: Its contents, its faults, their effects on evaluation, and its future use," *arXiv preprint arXiv:1306.1461*, 2013.

[28] Carol L Krumhansl and Petri Toiviainen, "Tonal cognition," *Annals of the New York Academy of Sciences*, vol. 930, no. 1, pp. 77–91, 2001.

[29] Vincent Lostanlen, Sripathi Sridhar, Brian McFee, Andrew Farnsworth, and Juan Pablo Bello, "Learning the helix topology of musical pitch," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 11–15.