STONE: SELF-SUPERVISED TONALITY ESTIMATOR

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

[1] Deezer Research, Paris, France

[2] Nantes Université, École Centrale Nantes, CNRS, LS2N, UMR 6004, F-44000 Nantes, France

ykong@deezer.com

ABSTRACT

Although deep neural networks can estimate the key of a musical piece, their supervision incurs a massive annotation effort. Against this shortcoming, we present STONE, the first self-supervised tonality estimator. The architecture behind STONE, named ChromaNet, is a convnet with octave equivalence which outputs a "key signature profile" (KSP) of 12 structured logits. First, we train ChromaNet to regress artificial pitch transpositions between any two unlabeled musical excerpts from the same audio track, as measured as cross-power spectral density (CPSD) within the circle of fifths (CoF). We observe that this self-supervised pretext task leads KSP to correlate with tonal key signature. Based on this observation, we extend STONE to output a structured KSP of 24 logits, and introduce supervision so as to disambiguate major versus minor keys sharing the same key signature. Applying different amounts of supervision yields semi-supervised and fully supervised tonality estimators: i.e., Semi-TONEs and Sup-TONEs. We evaluate these estimators on FMAK, a new dataset of 5489 real-world musical recordings with expert annotation of 24 major and minor keys. We find that Semi-TONE matches the classification accuracy of Sup-TONE with reduced supervision and outperforms it with equal supervision.

1. INTRODUCTION

Self-taught musicians can tell whether two pieces go "in tune" or "out of tune". To do so, they do not need to know the name of every key [1]. Meanwhile, in music information retrieval (MIR), current tonality estimators depend on a vocabulary of labels such as `C:maj` or `F:min`.

In this context, we aim to develop models which "learn by ear" like humans; i.e., from little or no annotated data. This goal is justified in practice by the fact that online digital music corpora are larger and more musically diverse than established MIR datasets, yet often lack expert metadata.

To overcome the need for large amount of labeled data, self-supervised learning (SSL) has emerged as an alternative paradigm to supervised learning, with numerous applications in speech and music processing [2–5].

The design of pretext tasks is a long-standing issue in SSL for audio. On one hand, some of them are meant as a pretraining step for general-purpose representation learning: contrastive predictive coding [6], deep metric learning [7], and self-distillation [8], to name a few. On the other hand, another family of pretext tasks is designed to suit a particular downstream task, such as tempo and pitch estimation [9, 10]. In this context, the concept of *equivariance* plays a central role. Loosely speaking, equivariance means that a certain parametric transformation of the input data forms a simple trajectory in the space of learned representations. Yet, equivariant SSL has never been used to study tonality, for lack of an adequate pretext task.

The main idea of our paper is that, even so *absolute* key labels are unknown, we can construct paired samples in which *relative* harmonic progressions serve as a learning signal for tonality estimation. Our contributions are:

STONE. To our knowledge, the first SSL framework whose model predictions correlates with key signatures. It comprises a new equivariant neural network named ChromaNet and a noncontrastive loss function based on cross-power spectral density (CPSD).

Downstream task. We extend STONE into Semi-TONE, a semi-supervised model that is tailored for 24-way key estimation. Semi-TONE performs on par with a supervised counterpart (Sup-TONE) while reducing dependency on annotated data by 90% [1].

FMAK. A new large dataset of 5489 real-world music recordings, collected from the Free Music Archive (FMA) and annotated by an expert for 24 major and minor keys, is available for free download [2].

2. RELATED WORK

2.1 Equivariant self-supervised learning in music

Equivariant SSL learns task-specific embeddings by representing the transformations which underlie its factors of

---

[1] Companion website: https://github.com/deezer/stone

[2] FMAK dataset (version 2): https://zenodo.org/records/12759100

variability: e.g., variations in pitch or tempo [10–12]. In particular, PESTO is a monophonic pitch estimator trained by learning the pitch shift of the same sample [9]. However, its extension to multipitch tracking is an open problem [13].

2.2 Computational models of tonality

Tonality estimation remains a relatively under-researched field, due to the scarcity of labeled data available for both training and evaluation purposes. The earliest methods were based on template matching [14–17]. Later, convolutional neural networks (convnets) and transformers appeared, treating the task as a supervised 24-class classification problem [18–20]. Among studies employing the same dataset, the convnet of Korzeniowski *et al.* [21] achieves the best performance. Against the lack of annotated data, prior work proposed to integrate key estimation with unsupervised autoencoding [18, 19]. However, their approach is computationally intensive and its evaluation is limited.

2.3 Annotated datasets for key estimation

Key detection datasets face a conundrum between diversity and reproducibility. The Million Song Dataset [22] offers key labels for diverse commercial music but lacks public audio. GiantSteps MTG Key (GSMK)[^3], GiantSteps Key (GSK) [23] and McGill Billboard datasets [24] offer public data, yet they are restricted in terms of genres and quantity. In particular, GSMK and GSK serve in the training and evaluation of supervised SOTA [21].

3. METHODS

Figure 1 illustrates our proposed method for STONE.

3.1 Artificial pitch transpositions of the CQT

We compute a constant-Q transform (CQT) with $Q = 12$ bins per octave and center frequencies ranging between $\xi_{\min} = 27.5$ Hz and $\xi_{\max} = 2^{99/12}\xi_{\min} = 8.37$ kHz.

Given a CQT matrix $\boldsymbol{x}$ and a integer $c \leq 15$, we reduce the number of rows from 99 down to 84 (7 octaves) by trimming the c lowest-frequency bins and $(15 - c)$ highest-frequency bins. This is tantamount by a pitch transposition by c semitones [9]. We denote the result by $T_c\boldsymbol{x}$ where $T_c\boldsymbol{x}[p,t] = \boldsymbol{x}[p - c, t]$ for each frequency $p < 84$ and time t.

3.2 ChromaNet: a convnet with octave equivalence

The cropped CQT matrix $T_c\boldsymbol{x}$ has a frequency range of $QJ = 84$ semitones with $Q = 12$ and $J = 7$ octaves. We define a 2-D fully convolutional network $f_{\boldsymbol{\theta}}$ with trainable parameters $\boldsymbol{\theta}$, operating on $T_c\boldsymbol{x}$ with no pooling over the frequency dimension. The last layer has a single channel and performs global average pooling over the time dimension.

The architecture $f_{\boldsymbol{\theta}}$ composes seven blocks, each of them composing a ConvNeXT block [25] and a time downsampling block, and layer normalization. It returns a vector in dimension QJ. While ConvNeXt blocks leaves the input

[^3]: https://github.com/GiantSteps/giantsteps-mtg-key-dataset
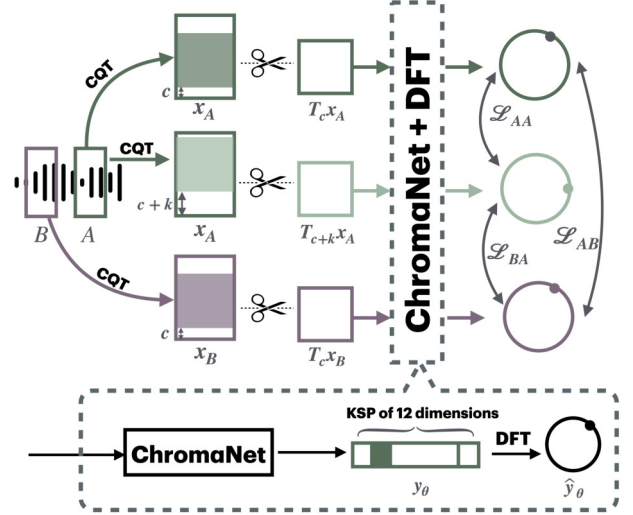

Figure 1. Overview of the equivariant pretext task in STONE. Given two segments A and B from an unlabeled musical recording, we compute their constant-Q transforms (CQT) and apply random crops by c and $(c + k)$ to simulate pitch transpositions. We feed them to ChromaNet, an equivariant neural network with octave equivalence, yielding a learned key signature profile (KSP) of 12 chromas. We compute the discrete Fourier transform (DFT) of each KSP and derive pairwise cross-power spectral densities (CPSD). Self-supervised losses $\mathscr{L}_{\mathrm{AA}}$, $\mathscr{L}_{\mathrm{AB}}$, and $\mathscr{L}_{\mathrm{BA}}$ are formulated as CPSD regression residuals in the complex domain.

resolution unchanged, time downsampling blocks decrease time resolution while preserving frequency resolution.

We compose the convnet $f_{\boldsymbol{\theta}}$ with a non-trainable operator g whose role is to guarantee octave equivalence. We roll the log-frequency axis into a spiral which makes a full turn at every octave, thus aligning coefficients in $f_{\boldsymbol{\theta}}(T_c\boldsymbol{x})$ of the form $(p \pm Qj)$ for integer j. The operator g sums these coefficients across octaves j for each pitch class q and applies a softmax transformation. We obtain a Q-dimensional vector $\boldsymbol{y}_{\boldsymbol{\theta}}$ whose entries are nonnegative and sum to one. We propose to call this vector a "key signature profile" (KSP):

$$\begin{aligned} \boldsymbol{y}_{\boldsymbol{\theta}}[q] &= (g \circ f_{\boldsymbol{\theta}})(T_c\boldsymbol{x})[q] \\ &= \frac{\exp\left(\sum_{j=0}^{J-1} f_{\boldsymbol{\theta}}(T_c\boldsymbol{x})[Qj+q]\right)}{\sum_{q'=0}^{Q-1}\exp\left(\sum_{j=0}^{J-1} f_{\boldsymbol{\theta}}(T_c\boldsymbol{x})[Qj+q']\right)}. \end{aligned} \quad (1)$$

For brevity, we do not recall the dependency of $\boldsymbol{y}_{\boldsymbol{\theta}}$ upon $\boldsymbol{x}$ nor c. Equation 1 resembles the extraction of chroma features [26], hence the proposed name of ChromaNet.

3.3 DFT over key signature profiles

With $Q = 12$, the discrete Fourier transform (DFT) of the KSP $\boldsymbol{y}_{\boldsymbol{\theta}}$ is

$$\widehat{\boldsymbol{y}}_{\boldsymbol{\theta}}[\omega] = \mathscr{F}\{\boldsymbol{y}_{\boldsymbol{\theta}}\}[\omega] = \sum_{q=0}^{11} \boldsymbol{y}_{\boldsymbol{\theta}}[q]e^{-2\pi i\omega q/12}, \quad (2)$$

where ω is an integer between 0 and 11 that is coprime to 12 for a full circular distribution over all 12 pitches. With

$\omega = 7$, a circular shift of $\mathbf{y_\theta}$ by seven chromas corresponds to a multiplication of $\widehat{\mathbf{y}}_\theta[\omega]$ by $e^{2\pi i 49/12} = e^{2\pi i/12}$. Hence, the phase of the complex number $\widehat{\mathbf{y}}_\theta[\omega]$ denotes a key modulation in the circle of fifths (CoF). Alternatively, $\omega = 1$ would correspond to a circle of semitones. Our paper evaluates both settings but only describes the CoF setting ($\omega = 7$) for the sake of conciseness.

3.4 Cross-power spectral density (CPSD)

Let us split the CQT matrix $\mathbf{x}$ into two disjoint time segments of equal length: $\mathbf{x} = (\mathbf{x}_A, \mathbf{x}_B)$. We denote the ChromaNet response for A by $\mathbf{y}_{\theta,A} = (g \circ f_\theta)(T_c \mathbf{x})$ idem for B. The circular cross-correlation between $\mathbf{y}_{\theta,A}$ and $\mathbf{y}_{\theta,B}$ is

$$\mathbf{R}_{\mathbf{y}_{\theta,A},\mathbf{y}_{\theta,B}}[k] = \sum_{q=0}^{Q-1} \mathbf{y}_{\theta,A}[q]\mathbf{y}_{\theta,B}[(q+k) \bmod Q] \quad (3)$$

for $0 \le k < 12$. Taking the DFT of the equation above yields the circular cross-power spectral density (CPSD)

$$\widehat{\mathbf{R}}_{\mathbf{y}_{\theta,A},\mathbf{y}_{\theta,B}}[\omega] = \mathscr{F}\{\mathbf{R}_{\mathbf{y}_{\theta,A},\mathbf{y}_{\theta,B}}\}[\omega] = \widehat{\mathbf{y}}_{\theta,A}[\omega]\widehat{\mathbf{y}}_{\theta,B}^*[\omega], \quad (4)$$

where the asterisk denotes complex conjugation.

3.5 Differentiable distance over the circle of fifths

Given a constant DFT frequency $\omega = 7$ and an arbitrary musical interval k in semitones, we compute the CPSD associated to the pair $(\mathbf{y}_{\theta,A}, \mathbf{y}_{\theta,B})$ and measure its half squared Euclidean distance to $e^{-2\pi i \omega k/Q}$ in the complex domain:

$$\mathscr{D}_{\theta,k}(\mathbf{x}_A, \mathbf{x}_B) = \frac{1}{2}\left|e^{-2\pi i \omega k/Q} - \widehat{\mathbf{R}}_{\mathbf{y}_{\theta,A},\mathbf{y}_{\theta,B}}[\omega]\right|^2. \quad (5)$$

Intuitively, in the case where $\widehat{\mathbf{y}}_{\theta,A}[\omega]$ and $\widehat{\mathbf{y}}_{\theta,B}[\omega]$ are both one hot-encoding of 12 dimensions, they will be mapped as complex numbers of module 1 on the border of the CoF, $\widehat{\mathbf{R}}_{\mathbf{y}_{\theta,A},\mathbf{y}_{\theta,B}}[\omega]$ measures the difference of phases on the CoF. Then, $\mathscr{D}_{\theta,k}(\mathbf{x}_A, \mathbf{x}_B)$ measures its deviation from the DFT basis vector $e^{-2\pi i \omega k/Q}$, which corresponds to the actual pitch shift k on the CoF. This distance is differentiable with respect to the weight vector θ.

3.6 Invariance loss

Although the contents of $\mathbf{x}_A$ versus $\mathbf{x}_B$ may differ in terms of melody, rhythm, and instrumentation, we assume them to be in the same key. This implies that ChromaNet responses $\mathbf{y}_{\theta,A}$ and $\mathbf{y}_{\theta,B}$ should be maximally correlated at the unison interval $k = 0$ and decorrelated for $k \ne 0$. In other words, the CPSD at the frequency ω should be maximal; i.e., equal to one. Thus, given an arbitrary pitch interval c, we define an *invariance loss* $\mathscr{L}_{AB}$, defined as the distance between $T_c \mathbf{x}_A$ and $T_c \mathbf{x}_B$ on the CoF. We obtain [4]:

$$\mathscr{L}_{AB}(\theta \,|\, \mathbf{x}, c) = \mathscr{D}_{\theta,0}(T_c \mathbf{x}_A, T_c \mathbf{x}_B) \quad (6)$$

3.7 Equivariance loss

We want the model f_θ to be equivariant to pitch transpositions. Hence, we define an *equivariance loss* $\mathscr{L}_{c,k}^{AA}$ as the distance between $T_c \mathbf{x}_A$ and $T_{(c+k)}\mathbf{x}_A$ on the CoF:

$$\mathscr{L}_{AA}(\theta \,|\, \mathbf{x}, c, k) = \mathscr{D}_{\theta,k}(T_c \mathbf{x}_A, T_{c+k}\mathbf{x}_A). \quad (7)$$

In theory, setting the architecture of f_θ to a ChromaNet should lead to equivariance by design, for any value of the weight vector θ. Yet, in practice, we observed that some values of θ break this property of equivariance, likely due to boundary artifacts in 2-D convolutions—a similar observation to PESTO [9]. For STONE, only minimizing the invariance loss $\mathscr{L}_c^{AB}$ causes the ChromaNet to collapse and predict a constant one-hot vector regardless of audio input $\mathbf{x}$ under certain hyperparameter choices, particularly for $\omega = 1$. To prevent this collapse, we penalize f_θ with the equivariance loss in Equation 7.

3.8 Combined invariance and equivariance loss

In addition, we penalize f_θ according to the following loss:

$$\mathscr{L}_{BA}(\theta \,|\, \mathbf{x}, c, k) = \mathscr{D}_{\theta,k}(T_c \mathbf{x}_B, T_{c+k}\mathbf{x}_A)., \quad (8)$$

i.e., the distance between ChromaNet responses $T_c \mathbf{x}_B$ and $T_{(c+k)}\mathbf{x}_A$ on the CoF. Observe that both these responses are already available after Equations 6 and 7. Therefore, the inclusion of Equation 8 in the loss comes at almost no extra computational cost during gradient backpropagation.

4. SELF-SUPERVISED KEY SIGNATURE PROFILES

4.1 Training on real-world unlabeled data

We collect 60k songs from the catalog of a music streaming service, with due permission. For each of them, we extract two disjoint segments $\mathbf{x}_A$ and $\mathbf{x}_B$ of duration equal to 15 seconds each. Following prior knowledge in music cognition [27], we set this duration to be as large as possible, considering the memory constraints of GPU hardware.

We implement ChromaNet and CPSD in PyTorch. The interval c (see Section 3.1) varies between zero and 15 semitones while the interval k (see Section 3.7) varies between -12 and 12 semitones. We define a CPSD-based stochastic loss function by combining Equations 6, 7, and 8:

$$\mathscr{L}^{CPSD}(\theta \,|\, \mathbf{x}, k, c) = \mathscr{L}_{AB}(\theta \,|\, \mathbf{x}, c)$$
$$+ \mathscr{L}_{AA}(\theta \,|\, \mathbf{x}, k, c) + \mathscr{L}_{BA}(\theta \,|\, \mathbf{x}, k, c), \quad (9)$$

where CQT samples $\mathbf{x}$ and intervals c and k are drawn independently and uniformly at random. We train the ChromaNet for 50 epochs using a cosine learning rate schedule with a linear warm-up. We use an AdamW optimizer with a learning rate of 10^{-3} and a batch size of 128.

The self-supervised procedure above learns an approximately equivariant mapping from CQT to key signature profiles (KSP). After training, we observe informally that for each input $\mathbf{x}$, most of the softmax activation in the KSP $\mathbf{y}_\theta$ is concentrated on a single pitch class. In other words, the loss $\mathscr{L}^{CPSD}$ (Equation 9) is sparsity-promoting.

4.2 Calibration on a C major scale

By learning to predict pitch transpositions between segments, STONE learns the notion of relative tonality, just like the relative pitch of musicians; however, it lacks a notion of absolute tonality. Thanks to the equivariant property of ChromaNet, we only need to introduce this notion via a single recording of a C major scale paired with C major chords. This calibration procedure resembles previous work in self-supervised fundamental frequency estimation [9, 11].

Denoting the C:maj calibration sample by $\boldsymbol{x}_{\text{cal}}$, we look up the index of its highest KSP coefficient in STONE:

$$q_{\text{cal}}(\boldsymbol{\theta}) = \arg\max_{0 \leq q' < Q} (g \circ f_{\boldsymbol{\theta}})(\boldsymbol{x}_{\text{cal}})[q']. \qquad (10)$$

Then, given a CQT matrix $\boldsymbol{x}$ from the test set, we realign its ChromaNet response $\boldsymbol{y} = (g \circ f_{\boldsymbol{\theta}})(\boldsymbol{x})$ via a pitch transposition of the learned KSP by $q_{\text{cal}}(\boldsymbol{\theta})$ semitones:

$$h_{\boldsymbol{\theta}}(\boldsymbol{y})[q] = \boldsymbol{y}[(q - q_{\text{cal}}(\boldsymbol{\theta})) \bmod Q]. \qquad (11)$$

4.3 Evaluation on real-world labeled data (FMAK)

FMAK is a subset of Free Music Archive dataset [28] containing 5489 songs that present a clear key and are in major or minor modes. We label these songs by ear. Each of the 24 keys is represented in FMAK by at least 89 songs. C:maj and A:min are the best represented, while G♯:maj and G♯:min are the least represented. Songs are distributed in major and minor modes evenly. Rock and electronic dance music are the best represented genres; jazz and blues are the least represented. To our knowledge, FMAK stands as the largest and most diverse MIR dataset with key annotation.

4.4 Key signature estimation accuracy (KSEA)

In compliance with MIREX [29], we propose the following figure of merit for key signature estimation:

$$\text{KSEA}(\boldsymbol{\theta}) = \frac{1}{N} \sum_{n=0}^{N-1} \Big(\boldsymbol{\delta}[s_n(\boldsymbol{\theta}) - S_n^{\text{ref}}]$$
$$+ \frac{1}{2} \boldsymbol{\delta}\big[||s_n(\boldsymbol{\theta}) - S_n^{\text{ref}}| - 6| - 1 \big] \Big), \qquad (12)$$

where $s_n(\boldsymbol{\theta}) = \arg\max_{0 \leq q < Q} (h_{\boldsymbol{\theta}} \circ g \circ f_{\boldsymbol{\theta}})(\boldsymbol{x}_n)[q]$ and $\boldsymbol{\delta}$ is the Kronecker symbol. 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.

4.5 Results on key signature estimation

We evaluate two variants of STONE on FMAK, all other things being equal: $\omega = 7$ (CoF) and $\omega = 1$. For comparison, we also evaluate the work of Korzeniowski *et al.*, the supervised state of of the art (SOTA) for this task [21], a convnet trained on GSMK. Lastly, we evaluate a feature engineering pipeline, requiring no supervision: i.e., we take the global average of the chromagram representation and extract the pitch class with highest energy.

Table 1 summarizes our results. We observe that, for both values of the CPSD frequency ω, STONE outperforms the feature engineering baseline. Furthermore, for $\omega = 1$, the KSEA approaches that of the supervised SOTA.

Table 1. Evaluation of self-supervised models on FMAK. KSEA denotes key signature estimation accuracy. We also report the supervised state of the art (SOTA) for comparison.

	Correct	Fifth	KSEA
Feature engineering	1599	981	38%
STONE ($\omega = 7$)	3587	1225	77%
STONE ($\omega = 1$)	**3883**	920	**79%**
Supervised SOTA [21]	4090	741	81%

Table 2. Ablation study of STONE ($\omega = 7$) on FMAK. CPSD denotes cross-power spectral density. KSEA denotes key signature estimation accuracy. We report a naive baseline (i.e., predict the key signature of C:maj and A:min for every sample) for comparison.

	Correct	Fifth	KSEA
STONE ($\omega = 7$)	**3587**	1225	**77%**
w/o octave equivalence	1052	1267	31%
w/o CPSD	1049	1267	31%
Baseline (predict C)	1049	1267	31%

4.6 Ablation study

The two main novel components of STONE are the ChromaNet (Section 3.2) on one hand and cross-power spectral density (CPSD) over learned key signature profiles (KSP) on the other hand. In order to evaluate their relative on performance, we conduct an ablation study: i.e., we substitute them by more conventional alternatives.

First, we replace the non-learned octave equivalence layer g (Equation 1) by a fully connected layer with same output size. Secondly, we replace the three CPSD-based losses (Equation 9) by 12-class cross-entropy losses. Intuitively, the first ablation disables equivariance in $f_{\boldsymbol{\theta}}$ while the second disables equivariance in $\mathcal{L}$. We observe that both ablations cause a collapse of SSL, leading it to predict the majority class (i.e., C:maj) on almost every sample. This suggests that both octave equivalence and CPSD are essential to the success of STONE.

5. SELF-SUPERVISED TONALITY ESTIMATION

5.1 Structured prediction

After having established that STONE learns to represent key signatures without any supervision, we turn to study its transferability to the well-known MIR problem of key estimation. For this purpose, we must accommodate the distinction between major and minor keys, thus doubling the output dimension of the ChromaNet from 12 to 24.

We note that key signature and mode are orthogonal concepts: each major key has exactly one relative minor and vice versa. These considerations suggest that the downstream task of key estimation may be formulated as structured prediction: i.e., in a 2-D label space. We encode structured labels in a matrix $\mathbf{Y}$ with 12 rows and two columns.
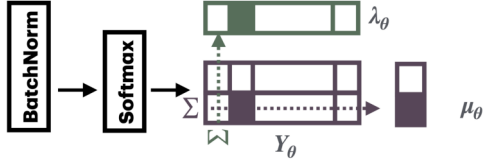
Figure 2. We modify the ChromaNet architecture of Figure 1 to accommodate structured prediction key signature and mode. We apply batch normalization per mode m and softmax over all coefficients, yielding a 12×2 matrix $\mathbf{Y}_{\boldsymbol{\theta}}(\boldsymbol{x})$. Summing $\mathbf{Y}_{\boldsymbol{\theta}}(\boldsymbol{x})$ over modes m yields a learned key signature profile $\lambda_{\boldsymbol{\theta}}(\boldsymbol{x})$ in dimension 12; summing $\mathbf{Y}_{\boldsymbol{\theta}}(\boldsymbol{x})$ over chromas q yields a pitch-invariant 2-dimensional vector $\mu_{\boldsymbol{\theta}}(\boldsymbol{x})$.

5.2 Batch normalization across key signatures

We modify the last layer of the ChromaNet $f_{\boldsymbol{\theta}}$ to output two channels instead of one. We also redefine the non-learnable operator g for octave equivalence to accommodate two channels, apply batch normalization with non-learnable parameters on each channel, and a softmax nonlinearity over all batch-normalized coefficients. This procedure normalizes each channel to null mean and unit variance over the training set, thus ensuring that both channels are activated and thus prevents a form of collapse during self-supervision.

The composition of g and $f_{\boldsymbol{\theta}}$, under their new definitions, yields a matrix $\mathbf{Y}_{\boldsymbol{\theta}}(\boldsymbol{x})$ with $Q = 12$ rows and two columns. By property of the softmax in g, all 24 coefficients in $\mathbf{Y}_{\boldsymbol{\theta}}(\boldsymbol{x})$ are positive and sum to one. As illustrated in Figure 2, we take advantage of this property to derive a key signature estimator $\lambda_{\boldsymbol{\theta}}$ and a mode estimator $\mu_{\boldsymbol{\theta}}$, respectively defined as row-wise and column-wise partial sums of $\mathbf{Y}_{\boldsymbol{\theta}}(\boldsymbol{x})$:

$$\lambda_{\boldsymbol{\theta}}(\boldsymbol{x})[q] = \sum_{m=0}^{1} \mathbf{Y}_{\boldsymbol{\theta}}(\boldsymbol{x})[q,m] \qquad (13)$$

$$\mu_{\boldsymbol{\theta}}(\boldsymbol{x})[m] = \sum_{q=0}^{11} \mathbf{Y}_{\boldsymbol{\theta}}(\boldsymbol{x})[q,m]. \qquad (14)$$

We verify that the 12-dimensional vector $\lambda_{\boldsymbol{\theta}}(\boldsymbol{x})$ is positive, sums to one, and is *equivariant* to pitch transpositions in $\mathbf{Y}_{\boldsymbol{\theta}}(\boldsymbol{x})$. Conversely, the 2-dimensional vector $\mu_{\boldsymbol{\theta}}(\boldsymbol{x})$ is positive, sums to one, and is *invariant* to pitch transpositions in $\mathbf{Y}_{\boldsymbol{\theta}}(\boldsymbol{x})$. We use $\lambda_{\boldsymbol{\theta}}$ as a substitute for $(g \circ f_{\boldsymbol{\theta}})$ in $\mathscr{L}^{\mathrm{CPSD}}$.

5.3 Self-supervised mode estimation

We now introduce a loss for self-supervised mode estimation. To this aim, we posit that mode is not only constant throughout the musical piece, but also remains invariant by pitch transposition. Therefore, going back to the notations from Section 4: $T_c\boldsymbol{x}_A$, $T_c\boldsymbol{x}_B$, and $T_{c+k}\boldsymbol{x}_B$ should elicit the same value of the mode estimator $\mu_{\boldsymbol{\theta}}$.

We recall the definition of binary cross-entropy (BCE) for 2-D vectors whose entries are positive and sum to one:

$$\mathrm{BCE}(\boldsymbol{\mu}, \boldsymbol{\mu}') = -\boldsymbol{\mu}[0] \log \boldsymbol{\mu}'[0] - \boldsymbol{\mu}[1] \log \boldsymbol{\mu}'[1]. \qquad (15)$$

We compute the pairwise BCE between mode estimator responses a ssociated to the three predictions of the self-supervised ChromaNet (see Figure 1)[5]:

$$\begin{aligned} \mathscr{L}^{\mathrm{BCE}}(\boldsymbol{\theta} \,|\, \boldsymbol{x}, c, k) = {} & \mathrm{BCE}(\mu_{\boldsymbol{\theta}}(T_c\boldsymbol{x}_B), \mu_{\boldsymbol{\theta}}(T_c\boldsymbol{x}_A)) \\ & + \mathrm{BCE}(\mu_{\boldsymbol{\theta}}(T_c\boldsymbol{x}_A), \mu_{\boldsymbol{\theta}}(T_{c+k}\boldsymbol{x}_A)) \\ & + \mathrm{BCE}(\mu_{\boldsymbol{\theta}}(T_c\boldsymbol{x}_B), \mu_{\boldsymbol{\theta}}(T_{c+k}\boldsymbol{x}_A)). \quad (16) \end{aligned}$$

We add the BCE-based loss in Equation 16 to the CPSD-based loss in Equation 9, thus yielding a full-fledged loss for self-supervised tonality estimation (STONE):

$$\mathscr{L}(\boldsymbol{\theta} \,|\, \boldsymbol{x}, c, k) = \mathscr{L}^{\mathrm{CPSD}}(\boldsymbol{\theta} \,|\, \boldsymbol{x}, c, k) + \mathscr{L}^{\mathrm{BCE}}(\boldsymbol{\theta} \,|\, \boldsymbol{x}, c, k). \qquad (17)$$

We train the modified ChromaNet to minimize $\mathscr{L}$ with the same optimization hyperparameters as in Section 4.1. The resulting model, named 24-STONE, performs key signature estimation with $\lambda_{\boldsymbol{\theta}}$ and mode estimation with $\mu_{\boldsymbol{\theta}}$. However these estimators are uncalibrated: i.e., $\lambda_{\boldsymbol{\theta}}$ only contains information of relative tonalities and $\mu_{\boldsymbol{\theta}}$ may swap relative major and minor. We calibrate them by means of a `C:maj` scale, via the same procedure as in Section 4.2.

5.4 Results on key and mode classification

We evaluate two variants of 24-STONE on FMAK: $\omega = 1$ and $\omega = 7$; as well as the supervised SOTA. We also evaluate the template matching algorithm of [17], requiring behavioral data but no supervision.

Table 3 summarizes our results. The 24-STONE model with $\omega = 7$ is best in the unsupervised category, although well below the supervised SOTA. However, setting ω to 1 dramatically hurts the MIREX score of 24-STONE, placing it below the naive baseline. Thus, formulating CPSD regression over the CoF (see Section 3.5) seems necessary for STONE to transfer to key and mode estimation, even so it is ouperformed by $\omega = 1$ in KSEA (see Section 4.4). With this result in mind, we set $\omega = 7$ in the rest of this paper.

6. SEMI-SUPERVISED TONALITY ESTIMATION

6.1 Supervising the ChromaNet

Thanks to structured prediction, the ChromaNet accommodates supervised training in the same label space as self-supervised training. Note that the 24-STONE losses $\mathscr{L}^{\mathrm{CPSD}}$ and $\mathscr{L}^{\mathrm{BCE}}$ involves pairwise comparisons between three items belonging to the same $\boldsymbol{x}$: i.e., two transposed versions of the same segment $\boldsymbol{x}_A$ and one from another segment $\boldsymbol{x}_B$. In this context, a simple way to introduce supervision is to replace the responses $\lambda_{\boldsymbol{\theta},B}$ and $\mu_{\boldsymbol{\theta},B}$ by "oracles" λ_{ref} and μ_{ref} which are informed by the ground truth.

Given the ground truth key signature q_{ref} and mode m_{ref}, one-hot encoding yields the sparse vectors $\lambda_{\mathrm{ref},c}(\boldsymbol{x})[q] = \boldsymbol{\delta}[(q - q_{\mathrm{ref}} - c) \bmod 12]$ and $\mu_{\mathrm{ref}}(\boldsymbol{x})[m] = \boldsymbol{\delta}[m - m_{\mathrm{ref}}]$, where c is a pitch interval in semitones (see Section 3.1). We use these oracles to write supervised variants of losses

---

[5] Compared to $\mathscr{L}^{\mathrm{CPSD}}$, we have swapped $\boldsymbol{x}_A$ with $\boldsymbol{x}_B$ in the first term. This is for compatibility with the supervised setting, as described in Section 6.1, so as to avoid an undefined BCE due to a logarithm of zero.

	Correct	Fifth	Relative	Parallel	Wrong	MIREX
Template matching [17]	2398	631	390	506	1564	53.4%
24-STONE ($\omega = 1$)	421	535	399	253	3881	15.6%
24-STONE ($\omega = 7$)	**2443**	628	1320	115	983	**57.9%**
Supervised SOTA [21]	3586	482	504	165	752	73.1%
Baseline (predict `C:maj`)	551	568	498	286	3586	19.0%

Table 3. Evaluation of self-supervised models for key and mode estimation on FMAK. We also report the supervised state of the art (SOTA) [21] and a naive baseline (i.e., predict `C:maj` for every sample) for comparison. See Section 5.4 for details.

$\mathscr{L}^{\text{CPSD}}$ and $\mathscr{L}^{\text{BCE}}$. This seamless switch from SSL to supervised learning requires no change of architecture nor optimizer. Thus, instead of using supervision as fine-tuning, we propose an alternated scheme: one epoch of SSL followed by one epoch of supervised learning, and so on.

6.2 Semi-TONE and Sup-TONE

Introducing supervision into 24-STONE yields a semi-supervised tonality estimator, or Semi-TONE for short. We alternate between self-supervised epochs on 60k unlabeled recordings (see Section 4.1) and supervised epochs on the 1159 songs in GSMK in which annotators agree. For the sake of comparison, we experiment with disabling SSL and training the ChromaNet directly on GSMK: hence a fully supervised tonality estimator, or Sup-TONE for short.

To compare their ability to learn from limited labeled data, we retrain Semi-TONE and Sup-TONE after subsampling GSMK at random by factors of 10 and 100.
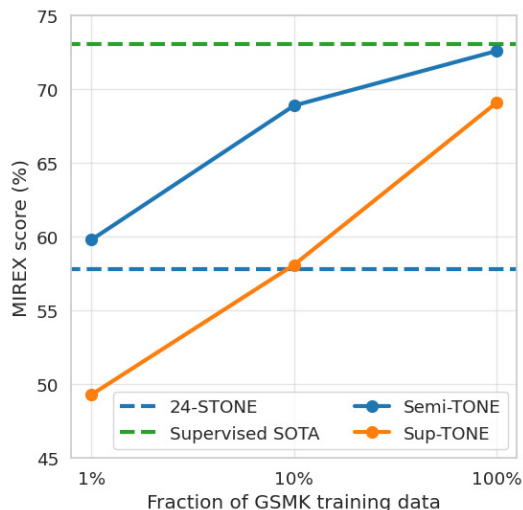

Figure 3. Evaluation of self-supervised (dashed blue), semi-supervised (solid blue), and supervised models (orange) on FMAK. All models use $\omega = 7$. We also report the supervised state of the art (SOTA) [21] in dashed green.

6.3 Results on key and mode classification

Figure 3 summarizes our results. We observe that Semi-TONE systematically outperforms Sup-TONE at any amount of training data. In particular, training Semi-TONE with 10% of GSMK leads to a comparable MIREX score as training Sup-TONE with 100% of GSMK. This result
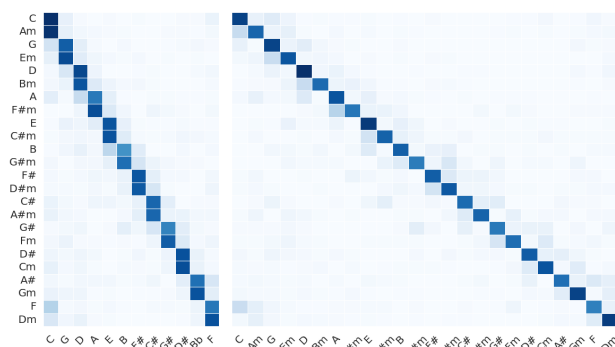

Figure 4. Confusion matrices of STONE (left, 12 classes) and Semi-TONE (right, 24 classes) on FMAK, both using $\omega = 7$. The axis correspond to model prediction and reference respectively, keys arranged by proximity in the CoF and relative modes. Deeper colors indicate more frequent occurences per relative occurence per reference key.

confirms the interest of our proposed pretext tasks towards the overarching goal of reducing human annotation effort.

Training Semi-TONE on the full GSMK dataset yields a MIREX score of 72.6%; i.e., roughly on par with the supervised SOTA (73.1%). Figure 4 shows the confusion matrix of calibrated STONE and Semi-TONE on FMAK. Although our methods do not outperform the SOTA on key estimation, it brings insights into a novel framework that does not require high supervision for training. Moreover, we note that self-supervision remains beneficial even when the full GSMK dataset is available for training. Therefore, a promising avenue of research is to scale up the dataset of unlabeled recordings (see Section 4.1), thus widening the gap between Semi-TONE and Sup-TONE on FMAK.

7. CONCLUSION

STONE learns key signature profiles (KSP) via equivariant self-supervised learning in the time–frequency domain. We have seen than a semi-supervised extension of STONE (semi-TONE) reduces expert annotation by 90% less at no loss of MIREX score compared to the fully supervised variant (sup-TONE). The primary limitation of our work resides in the inability of the STONE objective (CPSD, i.e., cross-power spectral density) to distinguish major keys from minor keys. Future work will study how STONE can be adapted to other pitch-relative MIR tasks.

8. REFERENCES

[1] D. Müllensiefen, B. Gingras, J. Musil, and L. Stewart, "The musicality of non-musicians: An index for assessing musical sophistication in the general population," *PLOS ONE*, vol. 9, no. 2, p. e89642, 2014.

[2] H. Zhu, Y. Niu, D. Fu, and H. Wang, "MusicBERT: A self-supervised learning of music representation," in *Proceedings of the ACM International Conference on Multimedia (MM)*, 2021, pp. 3955–3963.

[3] J. Spijkervet and J. A. Burgoyne, "Contrastive learning of musical representations," in *Proceedings of the International Society for Music Information Retrieval Late-Breaking/Demo Session (ISMIR-LBD)*, 2021.

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

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

[6] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "wav2vec: Unsupervised Pre-training for Speech Recognition," in *Proceedings of INTERSPEECH*, 2019.

[7] A. Saeed, D. Grangier, and N. Zeghidour, "Contrastive learning of general-purpose audio representations," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021.

[8] D. Niizumi, D. Takeuchi, Y. Ohishi, N. Harada, and K. 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.

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

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

[11] 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.

[12] G. Morais, M. E. Davies, M. Queiroz, and M. Fuentes, "Tempo vs. pitch: understanding self-supervised tempo estimation," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[13] F. Cwitkowitz and Z. Duan, "Toward fully self-supervised multi-pitch estimation," *arXiv preprint arXiv:2402.15569*, 2024.

[14] K. Noland and M. Sandler, "Signal processing parameters for tonality estimation," in *Proceedings of the Audio Engineering Society Convention (AES)*, 2007.

[15] S. Pauws, "Musical key extraction from audio," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, 2004.

[16] Á. Faraldo, E. Gómez, S. Jordà, and P. Herrera, "Key estimation in electronic dance music," in *Proceedings of the European Conference on Information Retrieval (ECIR)*, 2016.

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

[18] Y. Wu, E. Nakamura, and K. Yoshii, "A variational autoencoder for joint chord and key estimation from audio chromagrams," in *Proceedings of the Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2020.

[19] Y. Wu and K. Yoshii, "Joint chord and key estimation based on a hierarchical variational autoencoder with multi-task learning," *APSIPA Transactions on Signal and Information Processing*, 2022.

[20] H. Schreiber and M. Müller, "Musical tempo and key estimation using convolutional neural networks with directional filters," in *Proceedings of the International Sound and Music Computing Conference (SMC)*, 2019.

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

[22] T. Bertin-Mahieux, D. P. Ellis, B. Whitman, and P. Lamere, "The million song dataset," in *Proceedings of the International Society on Music Information Retrieval Conference (ISMIR)*, 2011.

[23] P. Knees, Á. Faraldo, P. Herrera, R. Vogl, S. Böck, F. Hörschläger, and M. L. Goff, "Two data sets for tempo estimation and key detection in electronic dance music annotated from user corrections," in *International Society for Music Information Retrieval Conference*, 2015. [Online]. Available: https://api.semanticscholar.org/CorpusID:15836728

[24] J. A. Burgoyne, J. Wild, and I. Fujinaga, "An expert ground truth set for audio chord recognition and music analysis," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, A. Klapuri and C. Leider, Eds. University of Miami, 2011.

[25] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A ConvNet for the 2020s," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, 2022, pp. 11 976–11 986.

[26] M. Müller, D. P. Ellis, A. Klapuri, and G. Richard, "Signal processing for music analysis," *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 6, pp. 1088–1110, 2011.

[27] P. Loui and D. Wessel, "Harmonic expectation and affect in Western music: Effects of attention and training," *Perception & Psychophysics*, vol. 69, pp. 1084–1092, 2007.

[28] M. Defferrard, K. Benzi, P. Vandergheynst, and X. Bresson, "FMA: A dataset for music analysis," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, 2017.

[29] 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 *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, 2014.