SEMI-SUPERVISED PIANO TRANSCRIPTION USING PSEUDO-LABELING TECHNIQUES

Sebastian Strahl, Meinard Müller
International Audio Laboratories Erlangen, Germany
{sebastian.strahl,meinard.mueller}@audiolabs-erlangen.de

ABSTRACT

Automatic piano transcription (APT) transforms piano recordings into symbolic note events. In recent years, APT has relied on supervised deep learning, which demands a large amount of labeled data that is often limited. This paper introduces a semi-supervised approach to APT, leveraging unlabeled data with techniques originally introduced in computer vision (CV): pseudo-labeling, consistency regularization, and distribution matching. The idea of pseudo-labeling is to use the current model for producing artificial labels for unlabeled data, and consistency regularization makes the model's predictions for unlabeled data robust to augmentations. Finally, distribution matching ensures that the pseudo-labels follow the same marginal distribution as the reference labels, adding an extra layer of robustness. Our method, tested on three piano datasets, shows improvements over purely supervised methods and performs comparably to existing semi-supervised approaches. Conceptually, this work illustrates that semi-supervised learning techniques from CV can be effectively transferred to the music domain, considerably reducing the dependence on large annotated datasets.

1. INTRODUCTION

Automatic music transcription (AMT) converts polyphonic music recordings into symbolic representations that encode which notes are played [1, 2]. The AMT output may be a MIDI-like transcription, containing for every note event information about the instrument, onset time, duration, and velocity. AMT is considered as one of the fundamental problems in music information retrieval (MIR) because its symbolic output can be used for subsequent tasks such as music synchronization, structure analysis, or cover song detection [3]. AMT is challenging since multiple instruments may be active at the same time, due to possible polyphonic activity per instrument, and because sound events may have overlapping harmonics [2].

Early approaches to AMT rely, e. g., on non-negative matrix factorization [4, 5], while most recent approaches use deep learning-based models [6–13]. The limiting factor in training neural networks for AMT, however, is the scarcity of labeled data. Creating such datasets typically requires manual labeling of each note present in a recording, which can be time-consuming, or relies on music synchronization techniques to align score information with recordings [11, 14]. The latter approach, however, may result in inaccurate labels due to issues such as playing errors or synchronization inaccuracies. Alternatively, one can create datasets with highly precise labels by utilizing instruments that allow automated playback or recording note activity. For instance, several piano datasets were automatically created using a Disklavier, which can synthesize MIDI files or log key activity during performance [15–17]. Since these piano datasets exist, many works [6–9, 12] focus on the special case of automatic piano transcription (APT). Still, it was observed that APT methods cannot generalize well across datasets due to overfitting [18].

In this work, we aim to improve model generalization of APT in scenarios with little labeled data by using semi-supervised learning (SSL), where the idea is to leverage unlabeled data during training. Unlabeled data can be obtained in large amounts as it does not depend on a labeling process. SSL has seen limited application in AMT, with Cheuk et al. [19] among the few to investigate this path. However, we argue that its full potential remains to be realized, especially when considering the significant achievements of SSL in computer vision (CV) [20, 21]. As our main contribution, we adapt techniques originally introduced in CV [22,23] to APT. More specifically, our method makes use of pseudo-labeling, consistency regularization, and distribution matching as outlined in the following.

In our approach, we use the extended Onsets and Frames model [7, 16], which jointly predicts onsets, offsets, frame activity, and velocities. The raw model outputs for onsets, offsets, and frames are each a piano roll-like representation that can be interpreted as probabilities per time–pitch bin. Initially, we pre-train this model in a supervised fashion using the available labeled data. Thereafter, the model is used to produce binary pseudo-labels for unlabeled data. Only sufficiently confident predictions are converted into pseudo-labels, i. e., those below the lower threshold are set to zero and those above the upper threshold are set to one, while the remaining predictions are considered as unreliable. Next, the model makes predictions for an augmented version of the same recording, where augmentation involves frequency masking [24] and addi-

tion of noise to the data. The predictions made for the augmented data are then used in combination with the pseudo-labels derived from the clean data to compute an additional unsupervised loss. Using an augmented version instead of a clean one encourages the model to produce consistent predictions under these kinds of augmentations and is thus called consistency regularization. As a third technique, we apply distribution matching, which ensures that the pseudo-labels follow the same marginal distribution as the reference labels, preventing the model from collapsing. To achieve this goal, we use an undersampling strategy. For reproducibility, we will provide our code [1] .

The rest of this paper is structured as follows: In Section 2, we give an overview of related work on AMT, SSL, and distribution matching in the context of pseudo-labeling. In Section 3, we describe all steps of the proposed approach. Section 4 describes our experimental setup as well as the experimental results. We conclude the paper in Section 5 with possible future research directions.

2. BACKGROUND AND RELATED WORK

2.1 Automatic Music Transcription

Most research on AMT is based on supervised learning. Sigtia et al. [6] proposed the the first end-to-end approach to APT. Hawthorne et al. [7] emphasized the importance of explicitly predicting onsets alongside frame activity, later extending their model in [16] to include explicit prediction of offsets. In [8], onset and offset estimation is formulated as a regression problem, which yields note predictions with improved temporal resolution. The attention-based Transformer architecture is used for APT [9, 12, 25] and multi-instrument AMT [10]. In [13], the Perceiver architecture is employed for multi-instrument AMT. Recently, AMT has been formulated as a conditional generative task: In [26], a diffusion model is trained to generate realistic piano rolls, being conditioned on the corresponding spectrograms.

Weakly supervised methods are proposed in [11], where unaligned pairs of scores and recordings are used for training, and in [27], where cross-version targets are used to replace pitch labels. Cheuk et al. [19] propose a semi-supervised approach to AMT, utilizing unlabeled data via virtual adversarial training (VAT). VAT [28] perturbs input data to induce substantial changes in the model's predictions and then encourages the model to produce consistent predictions under these perturbations. In [29], a fully self-supervised method is proposed for frame-level transcription. Their method encourages the concentration of energy around fundamental frequency candidates, invariance to timbral transformations, and equivariance to input translations in both time and frequency.

2.2 Semi-Supervised Learning

In SSL, the idea is to jointly learn from labeled and unlabeled data, and SSL is thus located between supervised and unsupervised learning [30,31]. The objective is to train a model that performs better than a reference model only trained on the labeled data using supervised learning. SSL has been successfully used in combination with deep learning, e. g., in CV [20, 21], for text classification [32], and also in MIR [33, 34]. For an overview of deep learning-based SSL methods, we refer to [20, 35]. Two important SSL paradigms relevant to this paper are pseudo-labeling and consistency regularization.

Pseudo-labeling, introduced in [36], uses the current classification model to produce artificial labels for unlabeled data. Continuing training with pseudo-labeled data encourages the model to make confident predictions for that data, effectively pushing decision boundaries away from the data points [35]. Maman and Bermano [11] already combined pseudo-labeling and weak supervision for AMT, but the pseudo-labels were updated only at the beginning of every expectation maximization iteration rather than being calculated on-the-fly as in [36].

Consistency regularization methods [37, 38] encourage that the model's predictions do not change if augmentations (e. g., random translation and addition of noise in the case of image classification [37, 38]) are applied to the unlabeled input data. In [37], this is achieved by adding a consistency loss term which penalizes disagreement in the predictions made for two augmented versions of the data.

The image classification method FixMatch [22] combines both pseudo-labeling and consistency regularization by using the current model to produce artificial labels given a weakly augmented input (e. g., horizontally flipped) to supervise the predictions made for a strongly augmented input (e. g., Cutout [39], where a randomly selected rectangular region is masked). In [40, 41], FixMatch proved to be effective for audio classification as well, where weak and strong augmentations were applied to spectrograms. FixMatch was also adapted to pixel-wise classification problems such as semantic image segmentation [42], which is similar to AMT from a technical point of view.

2.3 Distribution Matching

It is well-known that training classification models on class-imbalanced data is challenging because the models tend to be biased towards the majority classes [43]. Biased model predictions which do not follow a similar distribution as the reference labels are problematic for pseudo-labeling because the model may suffer from confirmation bias [44], where wrong predictions are reinforced. To avoid that problem, several approaches were proposed to match the class distribution of pseudo-labels with that of reference labels. Berthelot et al. [23] rescale the predicted class probabilities for unlabeled data in such a way that their marginal distribution is close to the marginal distribution of reference labels. Kim et al. [45] refine pseudo-labels by solving a convex optimization problem that aims to minimize the distance between pseudo-label distribution and reference label distribution while trying to preserve most information in the pseudo-labels. While Maman and Bermano [11] do not explicitly perform distribution matching for AMT, they set asymmetric thresholds for selecting pseudo-labels, increasing the impact of the minority class.

[1] https://github.com/groupmm/onsets_frames_semisup

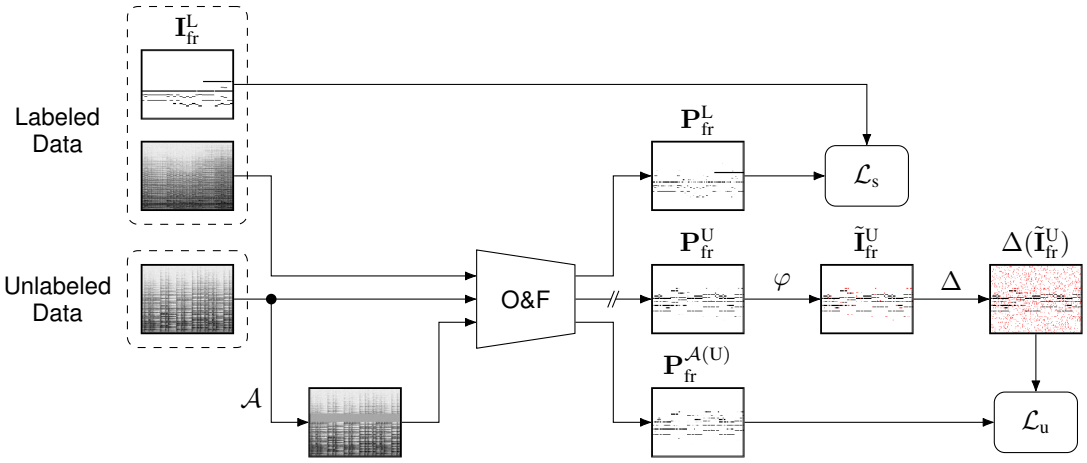

Figure 1: Detailed overview of our semi-supervised approach. The Onsets and Frames transcription model (O&F) [7, 16] is trained using both a supervised (upper branch) and an unsupervised loss (lower branches). Our method uses a clean version of unlabeled data to produce predictions, which, after thresholding (φ), are considered as pseudo-labels. Distribution matching (Δ) ensures that pseudo-labels and reference labels are similarly distributed. The pseudo-labels are used to supervise predictions made for an augmented ($\mathcal{A}$) version of the same data. The "interrupted" connection to the predictions made for the clean unlabeled input indicates that gradients are not backpropagated in this branch. For a better overview, we only show predictions, labels, and pseudo-labels for frame activity. Red color is used to represent NaN entries.

3. METHOD

In this section, we describe our proposed semi-supervised approach for learning APT. We first describe in Section 3.1 how the transcription model is trained in a supervised fashion. In Section 3.2, we explain how pseudo-labeling and consistency regularization can be used for semi-supervised training, and in Section 3.3, we explain the additional step of matching the pseudo-label distribution with the reference label distribution.

3.1 Supervised APT Baseline

We use the modified Onsets and Frames model [7, 16] and train our supervised APT baseline models similar to the original methodology. This model takes as input a log mel-scaled spectrogram with F frequency bins and T frames, and outputs onset, offset, frame activity, and velocity estimates. In this work, we focus on the involved classification problems and ignore velocity estimation for simplicity. Velocity estimation can be omitted without further consequences, as it is performed by an independent part of the model. We briefly explain how supervised learning is done using labeled data. The model outputs matrices $\mathbf{P}^{\mathrm{L}}_{\mathrm{on}}, \mathbf{P}^{\mathrm{L}}_{\mathrm{off}}, \mathbf{P}^{\mathrm{L}}_{\mathrm{fr}} \in [0,1]^{P\times T}$ for onset, offset, and frame activity, respectively. In this notation, P denotes the number of MIDI pitches considered, and the entries of the matrices represent probabilities of activities for all time–pitch bins. For instance, $\mathbf{P}^{\mathrm{L}}_{\mathrm{on}}(p,t)$ denotes the predicted probability of an onset with pitch p in frame t. The reference MIDI annotations with continuous-time note events are temporally quantized to match the input frame rate and converted into binary labels $\mathbf{I}^{\mathrm{L}}_{\mathrm{on}}, \mathbf{I}^{\mathrm{L}}_{\mathrm{off}}, \mathbf{I}^{\mathrm{L}}_{\mathrm{fr}} \in \{0,1\}^{P\times T}$, indicating bin-wise activities as described in [7, 16]. The supervised loss comprises three terms,

$$\mathcal{L}_{\mathrm{s}} = \lambda^{\mathrm{L}}_{\mathrm{on}}\mathcal{L}^{\mathrm{L}}_{\mathrm{on}} + \lambda^{\mathrm{L}}_{\mathrm{off}}\mathcal{L}^{\mathrm{L}}_{\mathrm{off}} + \lambda^{\mathrm{L}}_{\mathrm{fr}}\mathcal{L}^{\mathrm{L}}_{\mathrm{fr}}, \tag{1}$$

with the frame activity loss

$$\mathcal{L}^{\mathrm{L}}_{\mathrm{fr}} = \frac{1}{PT}\sum_{p=1}^{P}\sum_{t=1}^{T} \ell_{\mathrm{BCE}}(\mathbf{I}^{\mathrm{L}}_{\mathrm{fr}}(p,t), \mathbf{P}^{\mathrm{L}}_{\mathrm{fr}}(p,t)), \tag{2}$$

where ℓ_{BCE} denotes the binary cross entropy function and $\lambda^{\mathrm{L}}_{\mathrm{on}}, \lambda^{\mathrm{L}}_{\mathrm{off}}, \lambda^{\mathrm{L}}_{\mathrm{fr}} \in [0,1]$ are suitable loss weights. Onset and offset loss terms are defined analogously. Note that, in contrast to [7], we leave out the weighting of individual frames within the frame activity loss in Equation (2) for simplicity.

3.2 Pseudo-Labeling and Consistency Regularization

We now describe how our approach leverages unlabeled data, which is illustrated in Figure 1. Our method is mainly inspired by FixMatch [22], with the difference that we do not apply weak augmentations to produce pseudo-labels. Instead, we produce pseudo-labels using the unmodified, clean data, which has been found to yield nearly the same results in audio classification [40].

To obtain pseudo-labels for unlabeled data, we first compute the current model's predictions, $\mathbf{P}^{\mathrm{U}}_{\mathrm{on}}, \mathbf{P}^{\mathrm{U}}_{\mathrm{off}}, \mathbf{P}^{\mathrm{U}}_{\mathrm{fr}} \in [0,1]^{P\times T}$, given the clean version of the log mel-scaled spectrogram as input. For converting soft probabilities into binary pseudo-labels, we define a thresholding function

$$\varphi(x, \tau_{\mathrm{lo}}, \tau_{\mathrm{up}}) = \begin{cases} 1, & \text{if} \quad x \geq \tau_{\mathrm{up}}, \\ \mathrm{NaN}, & \text{if} \quad \tau_{\mathrm{lo}} < x < \tau_{\mathrm{up}}, \\ 0, & \text{if} \quad x \leq \tau_{\mathrm{lo}}, \end{cases} \tag{3}$$

where τ_{lo} and τ_{up} denote lower and upper threshold, respectively. We obtain the pseudo-labels $\tilde{\mathbf{I}}^{\mathrm{U}}_{\mathrm{on}}, \tilde{\mathbf{I}}^{\mathrm{U}}_{\mathrm{off}}$, and $\tilde{\mathbf{I}}^{\mathrm{U}}_{\mathrm{fr}}$ by elementwise application of the thresholding function to the model predictions, i. e.,

$$\tilde{\mathbf{I}}^{\mathrm{U}}_{\mathrm{fr}}(p,t) = \varphi(\mathbf{P}^{\mathrm{U}}_{\mathrm{fr}}(p,t), \tau_{\mathrm{lo}}, \tau_{\mathrm{up}}) \tag{4}$$

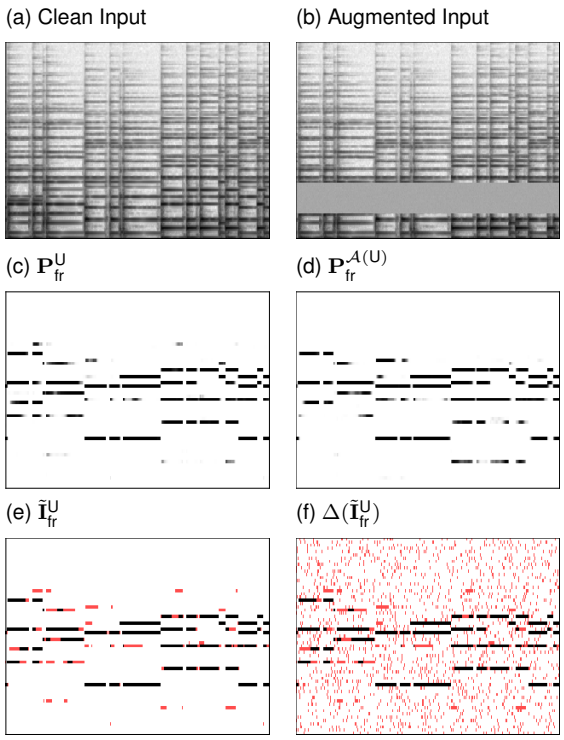

Figure 2: Examples of the representations involved in our semi-supervised method. Red color is used to represent NaN entries.

for $p \in [1:P]$, $t \in [1:T]$, and similarly for onsets and offsets. We use thresholds $\tau_{\text{lo}} = 0.05$ and $\tau_{\text{up}} = 0.95$ based on our observations in preliminary experiments, and we perform an ablation of this choice in Section 4. For illustration purposes, we refer to Figure 2, showing examples of clean model input, corresponding predictions $\mathbf{P}_{\text{fr}}^{\text{U}}$, and pseudo-labels $\tilde{\mathbf{I}}_{\text{fr}}^{\text{U}}$ in Figures 2a, 2c, and 2e, respectively, where NaN entries are represented by red color.

To perform consistency regularization, the pseudo-labels are used to supervise predictions made for an augmented version of the input. As in [40], we apply augmentations to the spectrograms. We opt for a simple augmentation pipeline which first applies frequency masking as described in [24], setting a randomly selected contiguous frequency band of up to 30 bins to the mean value of the spectrogram, and afterwards adds Gaussian noise with a standard deviation of 0.01 to the entire spectrogram. This choice of augmentation is inspired by the use of Cutout [39] in FixMatch [22] and the proposal of SpecAugment [24] as similar technique for spectrograms. We decided against temporal masking because this may completely remove information from the spectrogram regarding short events such as onsets. An example of such an augmented spectrogram is shown in Figure 2b. We denote the augmentation pipeline by $\mathcal{A}$, and the model's predictions for the augmented input are denoted by $\mathbf{P}_{\text{on}}^{\mathcal{A}(\text{U})}, \mathbf{P}_{\text{off}}^{\mathcal{A}(\text{U})}, \mathbf{P}_{\text{fr}}^{\mathcal{A}(\text{U})} \in [0,1]^{P \times T}$, respectively. An example of such predictions is shown in Figure 2d. Finally, the unsupervised loss is given by

$$\mathcal{L}_{\text{u}} = \lambda_{\text{on}}^{\text{U}} \mathcal{L}_{\text{on}}^{\text{U}} + \lambda_{\text{off}}^{\text{U}} \mathcal{L}_{\text{off}}^{\text{U}} + \lambda_{\text{fr}}^{\text{U}} \mathcal{L}_{\text{fr}}^{\text{U}}, \tag{5}$$

with the frame activity loss for unlabeled data,

$$\mathcal{L}_{\text{fr}}^{\text{U}} = \frac{1}{PT} \sum_{\substack{(p,t) \in [1:P] \times [1:T] \,: \\ \tilde{\mathbf{I}}_{\text{fr}}^{\text{U}}(p,t) \neq \text{NaN}}} \ell_{\text{BCE}}(\tilde{\mathbf{I}}_{\text{fr}}^{\text{U}}(p,t), \mathbf{P}_{\text{fr}}^{\mathcal{A}(\text{U})}(p,t)). \tag{6}$$

Onset and offset loss for unlabeled data are defined analogously. Only those time–pitch bins contribute to the loss, where the pseudo-labels have a value different from NaN. The loss is normalized by the total number of time–pitch bins for reducing the impact of the unsupervised loss if only a few predictions are confident. As for the supervised loss, we use suitable loss weights $\lambda_{\text{on}}^{\text{U}}, \lambda_{\text{off}}^{\text{U}}, \lambda_{\text{fr}}^{\text{U}} \in [0,1]$. Note that the gradient of $\mathcal{L}_{\text{u}}$ is not computed with respect to the predictions made for the clean version of the unlabeled input, which the "interrupted" connection in Figure 1 indicates. The overall loss function is obtained as the weighted sum of the supervised and the unsupervised loss,

$$\mathcal{L} = (1 - \lambda_{\text{u}}) \mathcal{L}_{\text{s}} + \lambda_{\text{u}} \mathcal{L}_{\text{u}}, \tag{7}$$

where $\lambda_{\text{u}} \in [0,1]$ controls the relative weighting of both terms. Following [7], we weight the individual terms in the supervised loss equally, i. e., $\lambda_{\text{on}}^{\text{L}} = \lambda_{\text{off}}^{\text{L}} = \lambda_{\text{fr}}^{\text{L}} = 1$. However, preliminary experiments suggested that better results may be achieved if the unsupervised offset loss is not used. Hence, our default setting is $\lambda_{\text{on}}^{\text{U}} = \lambda_{\text{fr}}^{\text{U}} = 1$ and $\lambda_{\text{off}}^{\text{U}} = 0$. The overall weight of the unsupervised loss is set to $\lambda_{\text{u}} = 0.05$. We explore the impact of these hyperparameter choices through ablation studies in Section 4.

3.3 Distribution Matching

The classification problems involved in training transcription models are heavily imbalanced because the labels typically have only a few non-zero entries. For example, the training set of the MAPS dataset [15] has labels, where only about 0.3% of all entries are ones for both onsets and offsets, and about 3.4% of all entries are ones for frame activity. Hence, the transcription model may be biased towards predicting zeros. To avoid model collapse, we apply distribution matching to the pseudo-labels.

In this paper, we employ a simple method to match the marginal pseudo-label distribution per mini-batch with that of the reference labels. The marginal distribution of the reference labels is estimated by counting zeros and ones across all training examples. These counting operations are denoted by Γ_0 and Γ_1. The following distribution matching method, explained using frame activity as an example, is similarly applied to onsets and offsets.

During training, we count the numbers of zeros and ones for every mini-batch of pseudo-labels, and will likely obtain a ratio $\Gamma_1(\tilde{\mathbf{I}}_{\text{fr}}^{\text{U}})/\Gamma_0(\tilde{\mathbf{I}}_{\text{fr}}^{\text{U}})$ that differs from the desired ratio $\Gamma_1(\mathbf{I}_{\text{fr}}^{\text{L}})/\Gamma_0(\mathbf{I}_{\text{fr}}^{\text{L}})$. The objective of the distribution matching operator, denoted by Δ, is to ensure that the ratio of zeros and ones is identical for reference labels and pseudo-labels, i. e.,

$$\frac{\Gamma_1(\mathbf{I}_{\text{fr}}^{\text{L}})}{\Gamma_0(\mathbf{I}_{\text{fr}}^{\text{L}})} = \frac{\Gamma_1(\Delta(\tilde{\mathbf{I}}_{\text{fr}}^{\text{U}}))}{\Gamma_0(\Delta(\tilde{\mathbf{I}}_{\text{fr}}^{\text{U}}))}. \tag{8}$$

	Thresholds		MAPS						MAESTRO						SMD					
			Note			Frame			Note			Frame			Note			Frame		
	τ_{on}	τ_{fr}	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Full																				
RV	0.50	0.50	80.9	70.6	75.1	85.9	72.0	77.9	-	-	-	-	-	-	-	-	-	-	-	-
OF	0.44	0.57	84.4	77.8	80.8	81.5	61.3	69.4	88.5	80.9	84.2	85.4	43.5	55.8	92.7	82.9	87.3	66.0	61.7	63.1
OF-SS4	0.35	0.34	84.7	79.6	81.9	78.3	67.5	72.0	93.3	82.7	87.5	84.5	53.2	63.5	94.7	85.5	89.7	63.1	69.0	65.2
Small																				
RV	0.50	0.50	86.2	57.1	68.2	90.0	43.9	58.2	-	-	-	-	-	-	-	-	-	-	-	-
OF	0.34	0.01	79.3	62.1	69.1	68.7	53.5	59.3	84.3	61.0	69.7	79.0	36.8	48.0	81.0	67.7	73.0	58.4	47.1	51.1
OF-SS4	0.05	0.01	78.2	75.9	76.7	62.3	69.9	65.0	93.8	78.4	85.0	73.6	56.3	61.8	93.6	80.2	85.9	52.3	68.1	58.2
One-Shot																				
RV	0.50	0.50	77.2	51.1	60.7	86.1	31.4	45.0	-	-	-	-	-	-	-	-	-	-	-	-
OF	0.02	0.01	66.5	56.0	60.2	67.4	35.3	45.2	76.5	50.9	59.9	76.7	23.3	34.0	69.0	57.9	62.1	56.3	32.3	39.8
OF-SS4	0.03	0.01	66.2	68.0	66.6	49.8	35.0	40.0	73.6	70.2	71.3	57.4	26.1	33.8	72.6	71.5	71.4	40.3	32.4	34.9
MB	0.50	0.50	88.2	86.5	87.3	84.4	76.7	79.6	92.6	87.2	89.7	77.4	76.1	76.0	-	-	-	-	-	-

Table 1: Performance metrics in percentages evaluated on the test sets of MAPS (ENSTDkAm and ENSTDkCl) and MAESTRO, and on the entire SMD dataset. Performance metrics are calculated per piece and then averaged over all pieces in the respective sets. As for the transcription models, RV is ReconVAT [19], OF is Onsets and Frames [16], OF-SS4 is our proposed semi-supervised method, and MB stands for Maman and Bermano [11]. Decision thresholds of OF and OF-SS4 are tuned using the group SptkBGAm of the MAPS dataset. F1 scores are highlighted in red for better readability.

To define Δ, we use undersampling as it is frequently used for class-imbalanced learning [46]. The distribution matching works as follows:

1. Determine whether the ratio $\Gamma_1(\tilde{\mathbf{I}}^{\text{U}}_{\text{fr}})/\Gamma_0(\tilde{\mathbf{I}}^{\text{U}}_{\text{fr}})$ is smaller or larger than the ratio $\Gamma_1(\mathbf{I}^{\text{L}}_{\text{on}})/\Gamma_0(\mathbf{I}^{\text{L}}_{\text{on}})$, i. e., whether there is an excess of zeros or ones, respectively, among the pseudo-labels.

2. Randomly select the required number of excess zeros or ones and convert them to NaN entries to obtain the desired ratio.

Distribution matching reduces the number of available pseudo-labels but ensures that the pseudo-labels within a mini-batch follow the same marginal distribution as the reference labels. An example of distribution-matched pseudo-labels is shown in Figure 2f.

4. EXPERIMENTS

4.1 Implementation Details

For our experiments, we use an open-source Pytorch implementation [2] of Onsets and Frames [7, 16]. Input representation and model architecture are unchanged compared to [7]. However, we do not ensure that input segments do not start in the middle of a note as it is done in [7]. We use a batch size of 8 each for labeled and unlabeled data and average losses across batches. We train our models using the Adam optimizer [47] with an initial learning rate of $6\text{e}{-5}$ and multiply the learning rate by a factor of 0.98 every 5k iterations. Also, we apply gradient clipping with norm 3. All audio recordings were downsampled to $16\,\text{kHz}$.

4.2 Datasets

We train and evaluate our models on three piano datasets: MAPS [15], MAESTRO V3.0.0 [16], and SMD [17].

[2] https://github.com/jongwook/onsets-and-frames

MAPS [15] contains isolated notes, chords, and complete piano pieces, but we only make use of the complete pieces. This dataset contains nine groups with 30 recordings each, where seven of the groups contain synthesized recordings, and the remaining two groups (ENSTDkAm and ENSTDkCl) contain real recordings which were automatically generated from MIDI files using a Disklavier. Following previous work [6, 7, 19], we use the groups with synthetic data as training data, and the real recordings as test data, and we remove the pieces from the training data which are also contained in the test data. This yields training and test sets of 139 and 60 recordings, respectively.

MAESTRO [16] and SMD [17] provide recordings together with the corresponding MIDI annotations automatically captured by a Disklavier. Both MAESTRO and SMD contain actual recordings of live performances, from the International Piano-e-Competition and played by music students, respectively. MAESTRO comprises 1276 performances, with the official data split assigning 962, 137, and 177 performances to the training, validation, and test set, respectively, and SMD comprises 50 performances.

4.3 Evaluation and Threshold Tuning

During inference, a decoding step is performed to obtain estimated note events from the network outputs [7, 16]. Two thresholds, τ_{on} and τ_{fr}, are applied to binarize onset and frame activity predictions. A note event is only recognized if an onset was detected, and the length of the note is determined based on the frame activity prediction. The offset prediction is not explicitly used during decoding.

Following existing literature, we evaluate model performance using note-based and frame-based metrics including precision (P), recall (R), and F1 score. Note-based metrics are computed using the *mir_eval* library [48], where a predicted note is considered as correct if its pitch matches that of a reference note and the onset is within $\pm 50\,\text{ms}$ of that reference note's onset.

Instead of using fixed thresholds τ_{on} and τ_{fr}, we tune these thresholds using a labeled validation set [27, 49]. We first determine an optimum τ_{on} via grid search so as to maximize the note F1 score, which does not depend on τ_{fr}. Since the frame-based metrics are computed based on the decoded note events, the frame F1 score is affected by both τ_{on} and τ_{fr}. We fix the previously found τ_{on} and determine the τ_{fr} that maximizes the frame F1 score.

4.4 Experimental Scenarios

To compare with [19], we adopt their three experimental scenarios which differ in the choice of the labeled data. The first scenario (***Full***) uses the full MAPS training set, the second scenario (***Small***) uses only the group `AkPnBcht` of the MAPS training set, which contains 23 non-overlapping piano pieces, and the third scenario (***One-Shot***) uses only a single recording (`chp_op31` from `AkPnBcht`) as labeled data. Note that for ***One-Shot***, the batch size for labeled data needs to be reduced to 1. In all scenarios, the MAESTRO training set is used as unlabeled data. We use the group `SptkBGAm` of the MAPS training set as validation data—which overlaps with the labeled training data in the ***Full*** scenario.

In all scenarios, we start training the transcription model from scratch following the training strategy described in Section 3.1 for 50k iterations, using only the labeled data and supervised learning. After that pre-training stage, we train for another 50k iterations using our proposed semi-supervised method as described in Section 3.2. We refer to this model as `OF-SS4`. For a fair supervised baseline in each scenario, we also continue training the pre-trained model for another 50k iterations on only the labeled data, which we will refer to as `OF`.

4.5 Main Results

The main results of our experiments are provided in Table 1, where the models of all scenarios are evaluated on the test sets of MAPS and MAESTRO, and also on the independent SMD dataset. First, we can observe that `OF-SS4` achieves better F1 scores than `OF` almost in all scenarios and across all datasets, with the frame F1 score in the ***One-Shot*** scenario being the exception. Most notably, `OF-SS4` achieves a note F1 score of 85.0 on the MAESTRO test set in the scenario ***Small***, which slightly exceeds the note F1 score 84.2 of `OF` in the scenario ***Full***. This shows that our semi-supervised approach is indeed effective, reducing the number of labeled performances by more than 80% for achieving comparable performance in this case. We further note that the optimum decision thresholds of `OF` and `OF-SS4` are extremely low for the scenarios ***Small*** and ***One-Shot***, indicating that threshold tuning is an important step if labeled training data is scarce.

For ReconVAT (`RV`) [19], we report for every scenario the performance of their semi-supervised method that achieved the highest note F1 score. Still, we observe that `OF-SS4` achieves higher note F1 scores than `RV` in all scenarios, e. g., 76.7 for `OF-SS4` compared to 68.2 for `RV` in the scenario ***Small***. Regarding the frame F1 score, no clear trend can be observed, with `OF-SS4` achieving a higher value for ***Small***, but lower values for ***Full*** and ***One-Shot***.

	τ_{lo}	τ_{up}	$\mathcal{A}$	Δ	$\lambda_{\text{off}}^{\text{U}}$	λ_{u}	N-F1	F-F1
`OF`	-	-	-	-	-	-	73.0	51.1
`OF-SS1`	0.05	0.95	-	-	0.0	0.05	0.1	3.0
`OF-SS2`	0.05	0.95	-	✓	0.0	0.05	82.4	9.4
`OF-SS3`	0.05	0.95	✓	-	0.0	0.05	82.7	57.6
`OF-SS4`	0.05	0.95	✓	✓	0.0	0.05	**85.9**	**58.2**
`OF-SS5`	0.25	0.75	✓	✓	0.0	0.05	74.6	51.6
`OF-SS6`	0.05	0.95	✓	✓	1.0	0.05	85.6	56.3
`OF-SS7`	0.05	0.95	✓	✓	0.0	0.01	72.8	51.5

Table 2: Results of an ablation study performed in the scenario ***Small***, evaluated on the independent SMD dataset [17]. N-F1 and F-F1 are note F1 score and frame F1 score in percentage, respectively.

As another reference, we include the weakly-supervised method by Maman and Bermano (`MB`) [11], which also relies on the Onsets and Frames transcription model [7, 16] but benefits from training on much more data and across various instrumentations. Our method does not reach the performance of `MB` in any scenario, but the performance gap is reasonably small given the difference in amount of training data, e. g., a note F1 score of 85.0 for `OF-SS4` in scenario ***Small*** compared to 89.7 for `MB` on MAESTRO.

4.6 Ablation Study

We perform an ablation study to evaluate the efficacy of the individual components of our semi-supervised method. The results of this study are shown in Table 2. The method `OF-SS1` performs pseudo-labeling without consistency regularization and distribution matching, where the performance metrics indicate potential model collapse. Better results are achieved when additionally using either distribution matching (`OF-SS2`) or consistency regularization (`OF-SS3`), achieving already better note F1 scores than the supervised baseline `OF`. The performance is further improved by combining both techniques, which results in our proposed method `OF-SS4`. The remaining ablations change the hyperparameter setting of our method, where less restrictive thresholds for selecting pseudo-labels (`OF-SS5`), calculating the unsupervised loss also for offsets (`OF-SS6`), or a reduced overall weight of the unsupervised loss (`OF-SS7`) yield worse results.

5. CONCLUSION

In this paper, we successfully transferred SSL techniques from CV to the MIR domain. More specifically, we applied pseudo-labeling, consistency regularization, and distribution matching for the task of APT, enabling the option to leverage unlabeled data during training. Thereby, the dependence on large annotated datasets is considerably reduced. For instance, using our semi-supervised approach, we observed reductions in the required amount of labeled data by up to 80% for achieving similar performance as a purely supervised baseline.

In future work, we plan to investigate other augmentation strategies, e. g., musically meaningful augmentations as in [18], to perform consistency regularization, and the extension of the method to the multi-instrument setting.

Acknowledgements: This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Grant No. 350953655 (MU 2686/11-2) and Grant No. 500643750 (MU 2686/15-1). The authors are with the International Audio Laboratories Erlangen, a joint institution of the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) and Fraunhofer Institute for Integrated Circuits IIS.

6. REFERENCES

[1] C. Raphael, "Automatic transcription of piano music," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, 2002, pp. 15–19.

[2] E. Benetos, S. Dixon, Z. Duan, and S. Ewert, "Automatic music transcription: An overview," *IEEE Signal Processing Magazine*, vol. 36, no. 1, pp. 20–30, 2019.

[3] M. Müller, *Fundamentals of Music Processing – Using Python and Jupyter Notebooks*, 2nd ed. Springer Verlag, 2021.

[4] P. Smaragdis and J. C. Brown, "Non-negative matrix factorization for polyphonic music transcription," in *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2003, pp. 177–180.

[5] E. Vincent, N. Bertin, and R. Badeau, "Adaptive harmonic spectral decomposition for multiple pitch estimation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 3, pp. 528–537, 2010.

[6] S. Sigtia, E. Benetos, and S. Dixon, "An end-to-end neural network for polyphonic piano music transcription," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 5, pp. 927–939, 2016.

[7] C. Hawthorne, E. Elsen, J. Song, A. Roberts, I. Simon, C. Raffel, J. H. Engel, S. Oore, and D. Eck, "Onsets and frames: Dual-objective piano transcription," in *Proceedings of the International Society for Music Information Retrieval Conference, (ISMIR)*, Paris, France, 2018, pp. 50–57.

[8] Q. Kong, B. Li, X. Song, Y. Wan, and Y. Wang, "High-resolution piano transcription with pedals by regressing onset and offset times," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3707–3717, 2021.

[9] C. Hawthorne, I. Simon, R. Swavely, E. Manilow, and J. H. Engel, "Sequence-to-sequence piano transcription with transformers," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, Online, 2021, pp. 246–253.

[10] J. Gardner, I. Simon, E. Manilow, C. Hawthorne, and J. H. Engel, "MT3: Multi-task multitrack music transcription," in *Proceedings of the International Conference on Learning Representations (ICLR)*, Virtual, 2022.

[11] B. Maman and A. H. Bermano, "Unaligned supervision for automatic music transcription in the wild," in *Proceedings of the International Conference on Machine Learning (ICML)*, Baltimore, Maryland, USA, 2022, pp. 14 918–14 934.

[12] K. Toyama, T. Akama, Y. Ikemiya, Y. Takida, W. Liao, and Y. Mitsufuji, "Automatic piano transcription with hierarchical frequency-time transformer," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, Milano, Italy, 2023, pp. 215–222.

[13] W. T. Lu, J. Wang, and Y. Hung, "Multitrack music transcription with a time-frequency perceiver," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Rhodes Island, Greece, 2023.

[14] C. Weiß, F. Zalkow, V. Arifi-Müller, M. Müller, H. V. Koops, A. Volk, and H. Grohganz, "Schubert Winterreise dataset: A multimodal scenario for music analysis," *ACM Journal on Computing and Cultural Heritage (JOCCH)*, vol. 14, no. 2, pp. 25:1–18, 2021.

[15] V. Emiya, R. Badeau, and B. David, "Multipitch estimation of piano sounds using a new probabilistic spectral smoothness principle," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 6, pp. 1643–1654, 2010.

[16] C. Hawthorne, A. Stasyuk, A. Roberts, I. Simon, C. A. Huang, S. Dieleman, E. Elsen, J. H. Engel, and D. Eck, "Enabling factorized piano music modeling and generation with the MAESTRO dataset," in *Proceedings of the International Conference on Learning Representations (ICLR)*, New Orleans, Louisiana, USA, 2019.

[17] M. Müller, V. Konz, W. Bogler, and V. Arifi-Müller, "Saarland music data (SMD)," in *Demos and Late Breaking News of the International Society for Music Information Retrieval Conference (ISMIR)*, Miami, Florida, USA, 2011.

[18] D. Edwards, S. Dixon, E. Benetos, A. Maezawa, and Y. Kusaka, "A data-driven analysis of robust automatic piano transcription," *IEEE Signal Processing Letters*, vol. 31, pp. 681–685, 2024.

[19] K. W. Cheuk, D. Herremans, and L. Su, "Reconvat: A semi-supervised automatic music transcription framework for low-resource real-world data," in *Proceedings of the ACM Multimedia Conference*, Virtual Event, China, 2021, pp. 3918–3926.

[20] X. Yang, Z. Song, I. King, and Z. Xu, "A survey on deep semi-supervised learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 9, pp. 8934–8954, 2023.

[21] A. Peláez-Vegas, P. Mesejo, and J. Luengo, "A survey on semi-supervised semantic segmentation," *CoRR*, vol. abs/2302.09899, 2023.

[22] K. Sohn, D. Berthelot, N. Carlini, Z. Zhang, H. Zhang, C. Raffel, E. D. Cubuk, A. Kurakin, and C. Li, "FixMatch: Simplifying semi-supervised learning with consistency and confidence," in *Advances in Neural Information Processing Systems (NeurIPS)*, Virtual, 2020.

[23] D. Berthelot, N. Carlini, E. D. Cubuk, A. Kurakin, K. Sohn, H. Zhang, and C. Raffel, "ReMixMatch: Semi-supervised learning with distribution matching and augmentation anchoring," in *Proceedings of the International Conference on Learning Representations (ICLR)*, Virtual, 2020.

[24] D. S. Park, W. Chan, Y. Zhang, C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "SpecAugment: A simple data augmentation method for automatic speech recognition," in *Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech)*, Graz, Austria, 2019, pp. 2613–2617.

[25] L. Ou, Z. Guo, E. Benetos, J. Han, and Y. Wang, "Exploring transformer's potential on automatic piano transcription," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Virtual and Singapore, 2022, pp. 776–780.

[26] K. W. Cheuk, R. Sawata, T. Uesaka, N. Murata, N. Takahashi, S. Takahashi, D. Herremans, and Y. Mitsufuji, "Diffroll: Diffusion-based generative music transcription with unsupervised pretraining capability," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Rhodes Island, Greece, 2023.

[27] M. Krause, S. Strahl, and M. Müller, "Weakly supervised multi-pitch estimation using cross-version alignment," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, Milano, Italy, 2023.

[28] T. Miyato, S. Maeda, M. Koyama, and S. Ishii, "Virtual adversarial training: A regularization method for supervised and semi-supervised learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 8, pp. 1979–1993, 2019.

[29] F. Cwitkowitz and Z. Duan, "Toward fully self-supervised multi-pitch estimation," *CoRR*, vol. abs/2402.15569, 2024.

[30] O. Chapelle, B. Schölkopf, and A. Zien, Eds., *Semi-Supervised Learning*. The MIT Press, 2006.

[31] J. E. van Engelen and H. H. Hoos, "A survey on semi-supervised learning," *Machine Learning*, vol. 109, no. 2, pp. 373–440, 2020.

[32] J. M. Duarte and L. Berton, "A review of semi-supervised learning for text classification," *Artificial Intelligence Review*, vol. 56, no. 9, pp. 9401–9469, 2023.

[33] S. Kum, J. Lin, L. Su, and J. Nam, "Semi-supervised learning using teacher-student models for vocal melody extraction," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, Montréal, Canada, 2020, pp. 93–100.

[34] M. Won, K. Choi, and X. Serra, "Semi-supervised music tagging transformer," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, Online, 2021, pp. 769–776.

[35] Y. Ouali, C. Hudelot, and M. Tami, "An overview of deep semi-supervised learning," *CoRR*, vol. abs/2006.05278, 2020.

[36] D.-H. Lee, "Pseudo-Label: The simple and efficient semi-supervised learning method for deep neural networks," in *Proceedings of the ICML 2013 Workshop on Challenges in Representation Learning (WREPL)*, Atlanta, GA, USA, 2013.

[37] S. Laine and T. Aila, "Temporal ensembling for semi-supervised learning," in *Proceedings of the International Conference on Learning Representations (ICLR)*, Toulon, France, 2017.

[38] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," in *Advances in Neural Information Processing Systems (NeurIPS)*, Long Beach, CA, USA, 2017, pp. 1195–1204.

[39] T. Devries and G. W. Taylor, "Improved regularization of convolutional neural networks with cutout," *CoRR*, vol. abs/1708.04552, 2017.

[40] S. Grollmisch and E. Cano, "Improving semi-supervised learning for audio classification with FixMatch," *Electronics*, vol. 10, no. 15, 2021.

[41] L. Cances, E. Labbé, and T. Pellegrini, "Comparison of semi-supervised deep learning algorithms for audio classification," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2022, no. 1, p. 23, 2022.

[42] M. M. i Rabadán, A. Pieropan, H. Azizpour, and A. Maki, "Dense FixMatch: A simple semi-supervised learning method for pixel-wise prediction tasks," in *Proceedings of the Northern Lights Deep Learning (NLDL) Workshop*, Tromsø, Norway, 2023.

[43] B. Krawczyk, "Learning from imbalanced data: Open challenges and future directions," *Progress in Artificial Intelligence*, vol. 5, no. 4, pp. 221–232, 2016.

[44] E. Arazo, D. Ortego, P. Albert, N. E. O'Connor, and K. McGuinness, "Pseudo-labeling and confirmation bias in deep semi-supervised learning," in *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*. Glasgow, United Kingdom: IEEE, 2020.

[45] J. Kim, Y. Hur, S. Park, E. Yang, S. J. Hwang, and J. Shin, "Distribution aligning refinery of pseudo-label for imbalanced semi-supervised learning," in *Advances in Neural Information Processing Systems (NeurIPS)*, Virtual, 2020.

[46] M. Buda, A. Maki, and M. A. Mazurowski, "A systematic study of the class imbalance problem in convolutional neural networks," *Neural Networks*, vol. 106, pp. 249–259, 2018.

[47] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proceedings of the International Conference for Learning Representations (ICLR)*, San Diego, California, USA, 2015.

[48] C. Raffel, B. McFee, E. J. Humphrey, J. Salamon, O. Nieto, D. Liang, and D. P. W. Ellis, "MIR_EVAL: A transparent implementation of common MIR metrics," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, Taipei, Taiwan, 2014, pp. 367–372.

[49] Y. Wu, B. Chen, and L. Su, "Polyphonic music transcription with semantic segmentation," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Brighton, UK, 2019, pp. 166–170.