COUNT THE NOTES: HISTOGRAM-BASED SUPERVISION FOR AUTOMATIC MUSIC TRANSCRIPTION

Jonathan Yaffe[1] **Ben Maman**[2] **Meinard Müller** [2] **Amit H. Bermano**[1]
[1] Tel Aviv University, Israel
[2] International Audio Laboratories Erlangen
jonathany@mail.tau.ac.il, ben.maman@audiolabs-erlangen.de

ABSTRACT

Automatic Music Transcription (AMT) converts audio recordings into symbolic musical representations. Training deep neural networks (DNNs) for AMT typically requires strongly aligned training pairs with precise frame-level annotations. Since creating such datasets is costly and impractical for many musical contexts, weakly aligned approaches using segment-level annotations have gained traction. However, existing methods often rely on Dynamic Time Warping (DTW) or soft alignment loss functions, both of which still require local semantic correspondences, making them error-prone and computationally expensive. In this article, we introduce CountEM, a novel AMT framework that eliminates the need for explicit local alignment by leveraging note event histograms as supervision, enabling lighter computations and greater flexibility. Using an Expectation-Maximization (EM) approach, CountEM iteratively refines predictions based solely on note occurrence counts, significantly reducing annotation efforts while maintaining high transcription accuracy. Experiments on piano, guitar, and multi-instrument datasets demonstrate that CountEM matches or surpasses existing weakly supervised methods, improving AMT's robustness, scalability, and efficiency. Our project page is available at `https://yoni-yaffe.github.io/count-the-notes`

1. INTRODUCTION

Automatic Music Transcription (AMT) converts audio recordings into symbolic, score-like representations. As a core task in Music Information Retrieval (MIR), AMT has applications in music education, analysis, production, and neural generation. However, it remains challenging, particularly for polyphonic and multi-instrument recordings, due to overlapping harmonics, complex timbres, and varying acoustic environments. Most AMT systems rely on strongly aligned training data, where each audio frame has an exact corresponding label [1–4]. While effective, creating such datasets is costly and labor-intensive, restricting AMT models to specific instruments, styles, and acoustic conditions. As an alternative, semi-supervised learning methods use weakly aligned segment-level annotations rather than frame-level labels, showing that imperfect supervision—such as unaligned transcriptions from different performances of the same piece, can still provide useful training targets [5–8].

One such method, NoteEM [5], applies an Expectation-Maximization (EM) framework to iteratively refine weak labels. Beginning with a transcriber trained on synthetic data, it alternates between aligning weak labels using the network's predicted features, and training the network with these labels. This strategy has achieved high transcription accuracy across diverse musical styles and instruments [5–7]. However, alignment methods like Dynamic Time Warping (DTW) [9] introduce synchronization errors, computational overhead, and label inconsistencies, even with improved neural features. This is especially true for note onset detection, where high temporal precision is crucial [1, 2, 5, 6]. Most critically, such approaches assume weak labels preserve event order, even if misaligned—an assumption that often fails in real-world scenarios, such as in arpeggios, where chords are performed as sequential notes.

As the main contribution of this article, we introduce **CountEM**, a novel AMT framework leveraging an even weaker form of supervision: note event counting, integrated with the Expectation-Maximization (EM) algorithm. Unlike supervised or weakly supervised methods that require structural alignment, CountEM uses note onset histograms to iteratively refine predictions and temporal estimates. A key insight of CountEM is that strict alignment steps based on approximate temporal ordering, enforced by methods like DTW, can be relaxed or eliminated. Instead of enforcing structure-preserving alignment, CountEM

Attribution: J. Yaffe, B. Maman, M. Müller, and A.H. Bermano, "Count The Notes: Histogram-Based Supervision for Automatic Music Transcription", in *Proc. of the 26th Int. Society for Music Information Retrieval Conf.,* Daejeon, South Korea, 2025.

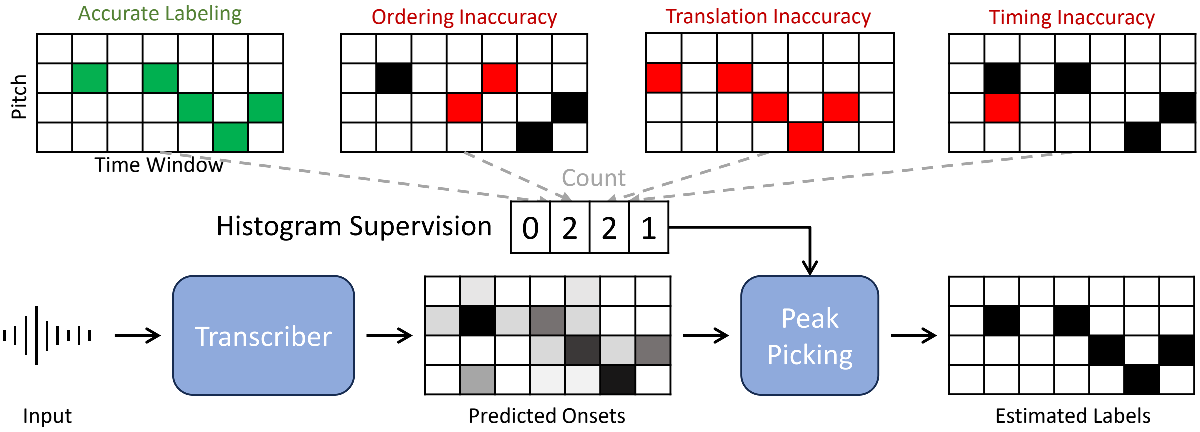

Figure 1. Estimating aligned labels from histograms by peak-picking. For each note in the histogram, the K most likely timings are selected according to the current predicted posteriorgram. Since misaligned labels reduce to the same histogram (top), possible timing inaccuracies common in weakly-aligned labels can be overcome.

counts note onsets within large time windows, using these counts alone as supervision. This reduces annotation effort while improving efficiency, flexibility, and robustness. Compared to DTW-based methods, histogram-based alignment is computationally simpler, and minimizes alignment errors caused by structural variations in musical performances.

To demonstrate the effectiveness of CountEM, we adapt the NoteEM framework to our histogram-based supervision approach. The model is initially trained on synthetic data, or other timing-accurate sources, before undergoing an iterative process of labeling and training. During labeling, the model generates onset estimates for each pitch over the prediction temporal window, and the K most probable timings are selected, where K is the supervised event count. See also Figure 1 for an illustration of the process. This method is applicable at various granularities, from entire audio tracks to smaller segments of 30 seconds, with longer windows providing weaker supervision. We evaluate CountEM on real-world datasets, showcasing its ability to generalize across diverse musical contexts, and demonstrate that it matches or surpasses existing weakly supervised methods. Even with large window sizes (up to entire tracks), it maintains high transcription accuracy. Furthermore, we demonstrate CountEM is robust to misalignments and annotation errors, enhancing AMT's scalability and extending its applicability to under-documented musical traditions.

The remainder of this article describes our approach in detail. Section 2 introduces the methods underlying CountEM, followed by Section 3, which presents the experiments and evaluation. Section 4 discusses key findings and implications, with directions for future research. Code and qualitative samples can be found on our project page. [1]

2. METHOD

Note histograms in musical performances can often be accurately derived from sheet music, particularly for Western classical music, which follows a musical score. CountEM leverages this information as coarse supervision for music transcription. The central insight of this work is that such counting supervision, which is easy to label and does not require precise timing or note ordering, can be a sufficient training signal. A second insight is that note onsets are prominent features in musical performances and remain consistent between a score and its rendition: If a note occurs K times in a musical score, then K onsets of that note will be perceived in an actual performance of that score. Indeed, studies on audio–score synchronization demonstrate improved alignment robustness when incorporating onset features [5, 10–12].

Other performance aspects, such as relative note timing, durations, intensity, and pitch fluctuations, vary by performer and interpretation. Traditional audio–score synchronization algorithms struggle with these variations, especially in polyphonic music, often leading to alignment errors [5, 6]. These errors stem from expressive timing and minor shifts in note order, such as in arpeggios. Effective alignment algorithms, especially for note onsets, must accommodate such variations. Recent transcription methods use DTW with neural onset features, followed by a refinement step that applies local temporal adjustments for each note independently [5, 6].

In contrast, our method alleviates the need for alignment and DTW by adopting a simpler, more flexible approach. Instead of enforcing strict temporal alignment, we use peak-picking to identify the K most probable onsets in a temporal window based on local maxima in the output signal. This straightforward, optimization-free process is robust to structural, timing, and ordering inaccuracies. The method follows

[1] `https://yoni-yaffe.github.io/count-the-notes`

Algorithm 1 CountEM

```
Input: audio $a_1, \dots a_N$, histog. $h_1, \dots h_N \in \mathbb{N}_0^P$
Output: model $f_\Theta$, labels $Y_1, \dots Y_N \in \{0,1\}^{T \times P}$
pre-train $f_\Theta$ (synthetic / other instrument)
$Y_i, d_i^{\text{hist}} = \text{None}, \infty \quad i = 1, \dots, N$
repeat
    for $i = 1$ to $N$ do
        $Y_i^{\text{temp}} = \text{PeakPick}(f_\Theta(a_i), h_i)$
        $h_i^{\text{pred}} = \Sigma_{t=1}^{T} f_\Theta(a_i)_t \in \mathbb{R}_+^P$
        $d_i^{\text{temp}} = \|h_i^{\text{pred}} - h_i\|_2$
        if $d_i^{\text{temp}} < d_i^{\text{hist}}$ then
            $Y_i, d_i^{\text{hist}} = Y_i^{\text{temp}}, d_i^{\text{temp}}$
        end if
    end for
    $\Theta = \underset{\Theta'}{\operatorname{argmin}} \sum_{i=1}^{N} \text{BCE}(f_{\Theta'}(a_i), Y_i)$
until $\sum_{i=1}^{N} d_i^{\text{hist}}$ converges
return $f_\Theta, Y_1, \dots Y_N$
```

an EM loop, alternating between label refinement and model improvement (see Section 2.1). The E-step refines labels using peak-picking (Section 2.2), while the M-step updates network parameters.

2.1 Expectation–Maximization

The EM process, outlined in Algorithm 1, consists of the following steps:

- **Initialization:** The model is pre-trained on fully supervised data from an easily accessible domain, such as synthetic data.

- **Expectation (E-step):** The model predicts a note onset posteriorgram (heatmap). The likelihoods in the posteriorgram are refined using top-K local-maxima peak picking for each pitch, based on its target number of occurrences, to estimate strongly-aligned onset labels. As a regularization, we only update the estimated label if the Euclidean distance between the current predicted histogram and the target histogram has improved.

- **Maximization (M-step):** We use the estimated strongly-aligned labels to update the model parameters using standard optimization [13].

The E- and M-steps are alternately repeated until convergence. We used 5 iterations for our experiments. The EM iterations progressively improve temporal localization without relying on detailed temporal annotation. Temporal precision is derived from the model itself, which is pre-trained on another domain.

2.2 Strong Alignment from Histograms

We use peak-picking to estimate precise time-aligned labels based on the target note histograms and the model's predictions. We assume a target histogram $h = (h_1, \dots, h_P)^\top \in \mathbb{N}_0^P$ where P is the number of considered pitches, and a predicted note onset posteriorgram $Z \in [0,1]^{T \times P}$, where T is the number of time frames. The posteriorgram Z can be interpreted as a predicted note onset heatmap, which we assume is computed as $Z = f_\Theta(a)$ for a given input audio representation a and a deep neural network f_Θ.

We assign an estimated label $Y \in [0,1]^{T \times P}$ using a peak-picking operator ("PeakPick" in Algorithm 1):

$$\Psi : [0,1]^{T \times P} \times \mathbb{N}_0^P \to \{0,1\}^{T \times P} \qquad (1)$$

which simply picks for each pitch $p \in \{1, \dots, P\}$ the K most likely temporal local peaks according to the predicted posteriorgram Z, where $K = h_p$ is the target number occurrences of the pitch according to the histogram. A position is considered to be a local peak if it is higher or equal to all its neighbors in a certain radius of frames, e.g., one frame.

Denoting $Y = \Psi(Z, h)$, for each pitch p the peak picker Ψ selects $K = h_p$ peaks from the p-th column of Z to define the p-th column of Y, where peak positions are binary-encoded (multi-hot). Note that by definition, it holds that $\Sigma_{t=1}^{T} Y_t = h \in \mathbb{N}_0^P$, i.e., the rows of Y sum up to the target histogram h.

2.3 Model Training

We experiment with two models: The Onsets and Frames architecture [1, 2] pre-trained on synthetic data [5], which we denote `Sy`, and the model of Kong et al. [3] pre-trained on the MAESTRO dataset, which we denote `Kg`. We optimize the mean binary cross-entropy (BCE) loss using an Adam optimizer [13]. To address the imbalance between positive and negative labels resulting from note onset sparsity, we assign a weight $w \geq 1$ to positive labels during training. This is done by applying a mask $M = w \cdot Y + (1 - Y)$ to the binary cross-entropy loss matrix, where Y is the estimated label. The loss function is computed as:

$$\mathcal{L}(f_\theta(a), Y) = \sum_{i,j} M_{i,j} \cdot \text{BCE}(f_\theta(a)_{i,j}, Y_{i,j}).$$

We set the weight w to 2 (`Sy`) or 1 (`Kg`), which from our observation provided approximately equal precision and recall. We apply pitch shift augmentation [5, 6, 14, 15], generating 11 pitch-shifted copies of the audio data, with shifts in the range of ± 5 semitones, and with an additional random fractional term in the range of ± 0.1 semitones to account for small tuning variation. Labels were computed only for the original copy and transposed accordingly for each augmented copy, enforcing pitch shift equivariance. All experiments were implemented in PyTorch and executed using two NVIDIA GeForce RTX 3090 GPUs. We used a batch size of 16 and trained models for 37.5K steps, except for Section 3.1, where we trained for 500K steps.

Model		Test			Train		
		P	R	F	P	R	F
Pre-trained Model							
	Sy	88.3	81.6	84.6	87.8	81.2	84.1
Histogram Supervision							
Rep. iter.	F/T	92.4	90.4	91.3	91.8	90.5	91.1
	180s	93.2	91.7	92.4	92.9	91.9	92.4
	120s	93.1	92.2	92.6	92.8	92.4	92.6
	60s	95.7	92.2	93.9	95.6	92.5	94.0
	30s	95.5	92.8	94.1	95.3	93.1	94.2
1-iter.	F/T	92.4	87.1	89.6	91.9	87.3	89.5
	60s	93.9	88.4	91.0	93.6	88.5	90.9
	Sup	98.7	93.1	95.8	98.8	93.4	96.0

Table 1. Note-level transcription results for training with histogram-based supervision on the MAESTRO dataset. We report Precision (P), Recall (R), and F-score (F) across different histogram window sizes (or Full Track). For reference, results include a baseline trained on synthetic data only (Sy) and a supervised model trained with ground-truth labels (Sup).

3. EXPERIMENTS

In this section, we present our experiments evaluating our approach across different datasets and instruments, including piano transcription and noisy histograms (Sections 3.1-3.2, MAESTRO dataset [2]), guitar transcription (Section 3.3, cross-dataset), and multi-instrument transcription including strings and winds (Section 3.4, cross-dataset). Evaluation metrics include note-level precision, recall, and F-score with a 50 ms onset tolerance.

3.1 Piano Transcription—MAESTRO Dataset

We first evaluate our method in a controlled setting using the MAESTRO dataset [2], which provides precise reference annotations generated automatically by a Disklavier. Instead of using these labels directly for training, we derive onset histograms by segmenting the audio and labels into smaller windows along the time axis, over which we compute histograms. These histograms serve as supervision for training, while evaluation is performed using the reference labels. To assess the impact of supervision levels, we test window lengths of 30 seconds, one minute, two minutes, three minutes, and entire tracks (up to 40 minutes).

Table 1 shows that our approach significantly improves transcription accuracy compared to the initial pre-trained model (Sy), even with full-track histograms (F/T), where F-score increases by over 6 % (from 84.6 to 91.3). Reducing the counting window further improves the F-score, as it better constrains onset timing, effectively increasing supervision. Performance approaches fully supervised levels for windows of one minute or less, indicating that the counting approach is effective even with temporally highly inaccurate labeling.

Model	Test			Train		
	P	R	F	P	R	F
Noisy Histogram Supervision						
60s0%	95.7	92.2	93.9	95.6	92.5	94.0
60s10%	93.1	92.0	92.5	93.0	92.2	92.6
60s20%	92.2	90.2	91.2	91.7	90.6	91.1
F/T0%	92.4	90.4	91.3	91.8	90.5	91.1
F/T10%	90.9	89.8	90.3	90.4	89.9	90.1
F/T20%	89.2	88.2	88.6	88.5	88.4	88.4

Table 2. CountEM robustness to noisy histograms on the MAESTRO dataset. We apply ±10% and ±20% random noise to simulate histogram errors and evaluate different window lengths as in Table 1.

We also observe that repeating the labeling process during training ("Rep. iter.") improves performance compared to training for the *same number of total steps* with a single labeling ("1-iter"), e.g., from 91.0 to 93.9 for one-minute windows.

3.2 Noisy Histograms

While fully supervised datasets like MAESTRO provide near-perfect histograms, labels for real-world recordings rely on musical scores, introducing potential discrepancies. For example, trills performed differently in audio and unaligned labels can cause minor inconsistencies. To assess the robustness of our approach, we train on the MAESTRO dataset with multiplicative random noise sampled from the uniform distribution $U[1-\alpha, 1+\alpha]$ at two levels ($\alpha \in \{0.1, 0.2\}$), introducing up to 10% and 20% noise. We conduct experiments using both one-minute and full-track histograms. Table 2 shows that while histogram errors slightly affect performance, the impact remains limited—no more than 3% even with 20% noise.

3.3 Guitar Transcription

As a next step, we evaluate our method on guitar datasets, namely GuitarSet [16] and the Guitar-Aligned Performance Scores (GAPS) dataset [7]. The annotation for GuitarSet was created by applying f_0 estimation on monophonic tracks obtained from hexaphonic pickup, followed by semi-automated methods for note onset and offset localization. The annotation for GAPS was done directly on polyphonic tracks by professional annotators, relying on recent neural network-based alignment techniques [7].

We compare two existing off-the-shelf models: The Onsets and Frames architecture [1, 2] pre-trained on synthetic data [5], and the model of Kong et al. [3] pre-trained on the MAESTRO dataset. We denote these models Sy and Kg, respectively. We train each of them using histogram supervision on each of the two datasets—GuitarSet and GAPS, which we denote

Gs and Gp, respectively. This yields four different configurations: SyGs, SyGp, KgGs, KgGp. We evaluate each configuration on each of the two datasets, enabling both intra- and inter-dataset (cross dataset) evaluation. We train each of the four configurations with histograms computed over different windows—one-minute windows (60s) and entire tracks (F/T).

The tracks in GuitarSet are all shorter than 30 seconds, therefore we only use entire-track histograms for it. Since GuitarSet is small (three hours) we train SyGs and KgGs on the entire set, however, only with histogram information. Therefore evaluation of SyGs and KgGs on GuitarSet measures the ability to restore the original time-aligned labels from the histogram information. When training on GAPS, we use the same train–test split as Riley et al. [7].

Results are shown in Table 3. It can be seen that our approach yields significant improvement over both baselines (Sy, Kg) of over 15% in F-score for both GuitarSet and GAPS, even when counting over entire tracks (SyGpF/T, KgGpF/T). For example, fine-tuning Sy on GAPS with histogram supervision over entire tracks (SyGpF/T) improves accuracy on GuitarSet from 66.2% to 84.6%.

When reducing the counting window on GAPS to one minute, Accuracy on GuitarSet slightly improves by 1.1% on average.

It can also be seen that by training on GuitarSet with only its histogram information we can restore its ground-truth strongly-aligned labels with accuracy of 88.9% (SyGsF/T) or 89.7% (KgGsF/T).

We further compare our results to previous work in weakly-supervised transcription. The model of Maman and Bermano [5] was fine-tuned from synthetic (the same pre-trained model we use) to self-collected guitar data. We denote this model by SySc. Accuracy of our model surpasses this model, improving on GuitarSet from 82.2% to 85.8% (SyGp60s) or 86.5% (KgGp60s), and on GAPS from 86.6% to 90% (SyGp60s) or 93% (KgGp60s).

The models of Riley et al. [7] were trained on GAPS with its time-aligned labels either from scratch ([7] Gp) or fine-tuned from piano ([7] KgGp). The labels were obtained by alignment of neural onset features, applying an initial DTW step, followed by a local-max refinement step for each note onset independently. Contrary to Riley et al. [7], we train on GAPS using histogram information only, i.e., weakening or completely omitting the DTW step. Our model's accuracy is slightly higher than [7] trained on GAPS from scratch, and slightly lower than [7] fined tuned on GAPS from MAESTRO, but on a comparable scale. This shows that the DTW step may be omitted with a small impact.

Most importantly, results show that our approach is robust across different architectures, and enables adaptation to guitar transcription from either synthetic (Sy) or piano (Kg) data pre-training.

Model	GuitarSet			GAPS		
	P	R	F	P	R	F
Pre-trained Models						
Sy	57.9	80.7	66.2	67.2	86.3	75.0
Kg	71.1	44.0	50.9	61.9	77.7	67.1
Histogram Supervision						
SyGsF/T	87.6	90.3	88.9	84.2	81.2	82.2
SyGpF/T	83.6	86.4	84.6	90.6	90.6	90.6
SyGp60s	85.6	86.5	85.8	89.8	90.1	90.0
KgGsF/T	89.3	90.1	89.7	85.4	89.0	87.1
KgGpF/T	83.6	88.1	85.5	93.3	92.5	92.9
KgGp60s	86.9	85.4	86.5	93.1	93.0	93.0
DTW + Refinement						
[7] Gp	92.4	81.8	86.1	94.9	92.1	93.4
[7] KgGp	91.1	85.9	88.1	95.0	93.6	94.3
[5] SySc	86.7	79.7	82.2	82.8	91.8	86.6

Table 3. Guitar transcription evaluation on the GuitarSet and GAPS datasets. We compare models pre-trained on synthetic data (Sy) and MAESTRO (Kg), trained on GuitarSet (Gs) and GAPS (Gp) using histograms from one-minute windows (60s) and entire tracks (F/T). See text for details.

3.4 Multi-Instrument Transcription

As a final, more challenging, and less controlled experiment, we evaluate the generalizability of the CountEM approach by applying our method to multi-instrument transcription using the MusicNet dataset [17], which features recordings of both solo and ensemble performances across various instruments. Unlike the MAESTRO dataset, MusicNet lacks full supervision, as its note labels were derived from aligning audio and MIDI files from different sources, introducing errors, particularly in onset timing [1,2,5]. However, a key advantage is that the musical structure was manually verified, ensuring consistency across performances. While fine-grained alignment remains imprecise, note histograms provide a stable and reliable signal, making this dataset well-suited for evaluating our histogram-based supervision approach in real-world, less curated conditions.

Another strength of MusicNet is its diversity in acoustics and instrumentation, making it well-suited for generalization across different musical contexts (zero-shot transcription).

We derive note histograms over entire tracks from unaligned labels. To obtain histograms over shorter chunks, we use loose alignment only to coherently subdivide audio and weakly-aligned labels. Minor errors in onset timing have little impact on histograms computed over 30- or 60-second windows. Future work could explore alternative segmentation techniques for further refinement.

Model	MAESTRO			GuitarSet			URMP			URMP (Histog.)		
	P	R	F	P	R	F	P	R	F	P	R	F
Pre-trained Model												
`Sy`	88.3	81.6	84.6	57.9	80.7	66.2	76.2	65.4	70.1	91.8	79.8	84.9
Histogram Supervision MusicNet Piano (ours)												
`30s`	93.0	88.2	90.4	77.8	82.5	79.4	70.1	79.6	74.5	80.1	90.8	85.0
`F/T`	92.1	85.8	88.7	81.2	80.1	79.8	77.3	75.1	76.1	89.7	87.1	88.3
Histogram Supervision MusicNet Full (ours)												
`32ms`	77.1	12.1	16.7	85.5	5.0	8.6	56.9	1.5	2.8	100.0	19.0	36.0
`100ms`	94.7	33.9	43.9	91.3	31.9	40.6	90.2	6.0	11.2	100.0	6.6	12.1
`500ms`	92.4	80.5	85.8	90.5	69.2	75.8	82.9	70.6	76.1	97.7	83.2	89.8
`30s`	94.5	86.0	89.9	88.5	75.4	80.3	82.2	79.9	80.9	93.0	90.4	91.6
`60s`	93.1	86.1	89.3	86.7	78.5	81.5	81.9	79.7	80.7	92.6	90.3	91.3
`F/T`	92.4	85.0	88.4	82.8	82.4	82.0	81.6	78.2	79.7	92.3	88.8	90.3
DTW + Refinement												
[5] `A1P1`	92.6	87.2	89.7	86.6	80.4	82.9	81.7	77.6	79.6	95.6	91.0	93.2
[5] `A1`	96.4	83.4	89.2	89.0	76.9	81.5	84.0	75.2	79.3	96.6	86.8	91.3

Table 4. Cross-dataset evaluation. Training was performed on MusicNet, with evaluation on MAESTRO, GuitarSet, and URMP. For URMP, we also report F-histogram, which does not enforce the 50ms onset threshold.

Note that while refined versions for the dataset exist [5], to demonstrate the efficacy of our approach we use the original, weakly-aligned labels.

We also note that we use the MusicNet dataset exclusively for training, as it lacks precise and reliable reference annotations. For evaluation, we again use the MAESTRO and GuitarSet datasets, along with the URMP dataset [18], which consists of string and wind instruments. In URMP the recordings are multi-tracked, where each track is monophonic, making annotations more accurate and reliable. While these labels are generally accurate, they are not perfectly precise [4]. To account for potential timing inaccuracies, we report both the standard 50ms onset F-score, and a high-tolerance metric, referred to as onset *F-histogram*. It is computed similarly to the F-score, but without the 50ms threshold. It compares the sets without considering timing, and serves as an upper bound in cases of annotation errors in onset timing.

We experimented with both pre-trained models appearing in Section 3.3, however, the synthetic pre-trained model (`Sy`) performed better than the piano pre-trained one (`Kg`). We postulate this is thanks to the diversity in the data used to pre-train `Sy` (despite being synthetic). Therefore, presented results are from `Sy`, also used by Maman and Bermano [5], but fine-tuned on MusicNet with our approach.

As shown in Table 4, our approach improves over the synthetic baseline, even with full-track histograms (`F/T`), increasing accuracy on MAESTRO from 84.6% to 88.7%, and reaching 90.4% for half-minute segments. It slightly outperforms a model from previous work trained with alignment and pseudo-labels ([5] `A1P1`) while relying on a much simpler label estimation method. Notably, our results even with full-track histograms match results using DTW and local-max refinement ([5] `A1`), suggesting that DTW may not be essential for this task.

Lastly, we note that when reducing the window size below `100ms`, accuracy drastically drops, contrary to the MAESTRO dataset where a single frame (corresponding to full supervision) provides best results. This demonstrates that the MusicNet labels contain errors in onset timing, and also shows that our approach can overcome them, as illustrated in Figure 1.

4. CONCLUSION

In this work, we introduced CountEM, a novel framework for AMT that leverages histogram-based supervision to eliminate the need for explicit temporal alignment. By replacing traditional alignment strategies with a simple peak-picking mechanism, CountEM reduces computational overhead while improving flexibility. Extensive experiments across piano, guitar, and multi-instrument datasets demonstrated its robustness, achieving performance comparable to or surpassing existing weakly-supervised methods with a significantly simplified label estimation process.

Looking ahead, CountEM's principles could extend to tasks such as instrument recognition, rhythm analysis, and lyrics transcription, particularly in complex polyphonic settings. Further exploration of weakly- and semi-supervised learning strategies could enhance transcription accuracy while minimizing annotation costs. By shifting towards more efficient and scalable supervision mechanisms, CountEM paves the way for data-efficient approaches to music transcription across diverse musical contexts.

5. ACKNOWLEDGEMENTS

This research was supported in part by the Len Blavatnik and the Blavatnik family foundation and ISF grant number 1337/22. This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Grant No. 500643750 (MU 2686/15-1). The International Audio Laboratories Erlangen are a joint institution of the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) and Fraunhofer Institute for Integrated Circuits IIS.

6. REFERENCES

[1] 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.

[2] 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. [Online]. Available: https://openreview.net/forum?id=r1lYRjC9F7

[3] 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 of Audio, Speech, and Language Processing*, vol. 29, pp. 3707–3717, 2021. [Online]. Available: https://doi.org/10.1109/TASLP.2021.3121991

[4] 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.

[5] 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.

[6] X. Riley, D. Edwards, and S. Dixon, "High resolution guitar transcription via domain adaptation," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Seoul, South Korea, 2024, pp. 1051–1055.

[7] X. Riley, Z. Guo, and S. Edwards, Drew abd Dixon, "Gaps: A large and diverse classical guitar dataset and benchmark transcription model," *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR), San Francisco, USA*, 2024.

[8] F. Zalkow and M. Müller, "CTC-based learning of chroma features for score-audio music retrieval," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 2957–2971, 2021.

[9] M. Müller, *Information Retrieval for Music and Motion*. Springer Verlag, 2007.

[10] S. Ewert, M. Müller, and P. Grosche, "High resolution audio synchronization using chroma onset features," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Taipei, Taiwan, 2009, pp. 1869–1872.

[11] Y. Özer, M. Istvanek, V. Arifi-Müller, and M. Müller, "Using activation functions for improving measure-level audio synchronization," in *Proceedings of the 23rd International Society for Music Information Retrieval Conference, ISMIR 2022, Bengaluru, India, December 4-8, 2022*, P. Rao, H. A. Murthy, A. Srinivasamurthy, R. M. Bittner, R. C. Repetto, M. Goto, X. Serra, and M. Miron, Eds., 2022, pp. 749–756. [Online]. Available: https://archives.ismir.net/ismir2022/paper/000090.pdf

[12] J. Zeitler, B. Maman, and M. Müller, "Robust and accurate audio synchronization using raw features from transcription models," *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR), San Francisco, USA*, 2024.

[13] 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. [Online]. Available: http://arxiv.org/abs/1412.6980

[14] J. Thickstun, Z. Harchaoui, D. P. Foster, and S. M. Kakade, "Invariances and data augmentation for supervised music transcription," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, Canada, 2018, pp. 2241–2245.

[15] A. Riou, S. Lattner, G. Hadjeres, and G. Peeters, "PESTO: Pitch estimation with self-supervised transposition-equivariant objective," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, Milano, Italy, 2023, pp. 535–544.

[16] Q. Xi, R. M. Bittner, J. Pauwels, X. Ye, and J. P. Bello, “Guitarset: A dataset for guitar transcription,” in *Proceedings of the 19th International Society for Music Information Retrieval Conference, ISMIR 2018, Paris, France, September 23-27, 2018*, E. Gómez, X. Hu, E. Humphrey, and E. Benetos, Eds., 2018, pp. 453–460. [Online]. Available: http://ismir2018.ircam.fr/doc/pdfs/188_Paper.pdf

[17] J. Thickstun, Z. Harchaoui, and S. M. Kakade, “Learning features of music from scratch,” in *Proceedings of the International Conference on Learning Representations (ICLR)*, Toulon, France, 2017. [Online]. Available: https://openreview.net/forum?id=rkFBJv9gg

[18] B. Li, X. Liu, K. Dinesh, Z. Duan, and G. Sharma, “Creating a multitrack classical music performance dataset for multimodal music analysis: Challenges, insights, and applications,” *IEEE Transactions on Multimedia*, vol. 21, no. 2, pp. 522–535, 2019.