Harmonic-Aware Frequency and Time Attention for Automatic Piano Transcription

Qi Wang, *Member, IEEE*, Mingkuan Liu, Changchun Bao, *Senior Member, IEEE*, and Maoshen Jia, *Senior Member, IEEE*

***Abstract*—Automatic music transcription (AMT) is to transcribe music audio into note symbol representations. Concurrent notes overlapping in the frequency and time domains still hinder the performance of polyphonic piano transcription in current studies. In this work, we develop an attention-based method for piano transcription, where we propose a harmonic-aware attention to capture the musical frequency structure, and a local time attention to model temporal dependencies. The harmonic-aware frequency attention not only emphasizes the relationship between the obvious harmonics, but also extracts the correlation in the residual non-harmonic component. The time attention mechanism is improved using the learnable attention range masks to model frame-wise short-term dependencies on different subtasks. Experiments on the MAESTRO dataset demonstrate that the proposed system achieves state-of-the-art transcription performance on both frame-wise and note-wise F1 metrics. Considering the influence of the piano pedals' dynamic behavior on note duration, a note duration modification method is also proposed. With a more accurate annotation of the offset on MAESTRO, the transcription performance is further improved.**

***Index Terms*—Piano transcription, frequency attention, time attention, harmonic mask, piano pedal.**

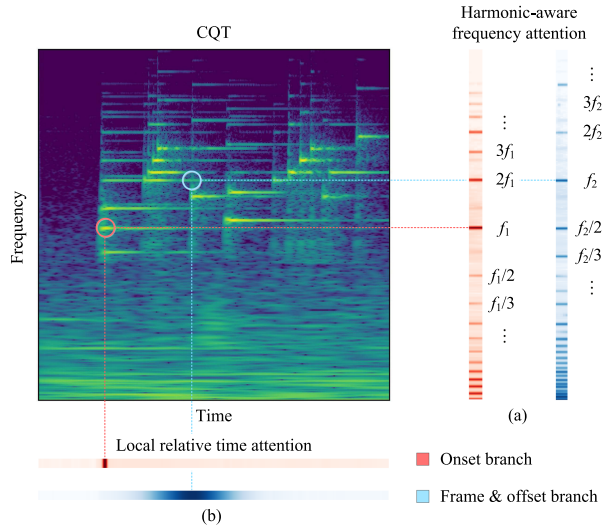

Fig. 1. (a) Harmonic-aware frequency attention aggregates the structured information in the spectrum. (b) Local relative time attention models the temporal dependencies on different branches.

I. Introduction

AUTOMATIC music transcription (AMT) [1] aims to transform music audio signals to symbol representations, such as piano rolls, MIDI files and sheet music. The AMT has a broad range of applications in music information retrieval (MIR), including automatic music generation, music search, music education, and musicology research [2].

Piano transcription is a widely studied task in AMT. As the piano music is highly polyphonic, its transcription is challenging. To determine the presence of the pitch in audio, many discriminative models were used in the early transcription systems, such as support vector machines (SVM) [3], [4]. Non-negative matrix factorization (NMF) and probabilistic latent component analysis (PLCA) are also popular methods, which can decompose polyphonic piano music spectra into multiple note spectrum templates and activations [5], [6], [7], [8], [9].

Manuscript received 22 January 2024; revised 13 May 2024 and 14 June 2024; accepted 15 June 2024. Date of publication 28 June 2024; date of current version 26 July 2024. This work was supported by the National Natural Science Foundation of China under Grant 62001012. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Emmanouil Benetos. *(Corresponding author: Changchun Bao.)*

The authors are with the Faculty of Information Technology, Beijing University of Technology, Beijing 100021, China (e-mail: wangqi91@bjut.edu.cn; liumkuan@emails.bjut.edu.cn; chchbao@bjut.edu.cn; jiamaoshen@bjut.edu.cn).

Digital Object Identifier 10.1109/TASLP.2024.3419441

With recent advancements in deep learning, the neural networks (NNs) have been widely used in AMT methods. Convolutional neural networks (CNNs) were employed to extract the acoustic features by taking a two-dimensional spectrogram as an input [10]. Recurrent neural networks (RNNs) were also proposed to learn regressions from time-series audio to the labelled pitches [11]. In addition to the acoustic model of CNNs, RNNs were used to model the music language information [12].

The multi-task learning has been proven effective in piano transcription by predicting multiple targets, including pitch, onset, offset and velocity. Hawthorne et al. designed a dual-objective onsets and frames (OAF) system, which used CNNs followed by a bidirectional RNN to estimate frame-wise pitches and onsets jointly [13]. Then the OAF system was improved using an adversarial learning method [14]. Kelz et al. proposed a transcription model with an extra note offset detection branch, and investigated the impact of additional prediction targets [15]. For the subtask of onset and offset detection, Kong et al. designed high-resolution note targets by regressing precise onset and offset times [16].

With the success of the attention mechanism in many machine learning fields, additive attention was also introduced into the OAF system to explore the temporal correlation in each branch [17]. Hawthorne et al. [18] and Gardner et al. [19]

directly utilized a generic Transformer for end-to-end piano transcription. To explore the potential of Transformer in the multi-task framework, Ou et al. applied the Transformer to each subtask, and only the note velocity estimation performance was improved [20]. In [21], a two-level hierarchical frequency-time Transformer was applied to the multi-task framework, achieving state-of-the-art piano transcription results. Beyond the piano transcription, Lu et al. proposed the SpecTNT, which models both spectral and temporal sequences on the time-frequency representation [22]. Similarly, the Perceiver TF models the time-frequency representation of audio input for multitrack transcription [23]. Taking the spectral and temporal dependencies into account has been proven effective in music transcription.

The harmonic structure in the frequency domain is crucial for pitch estimation. Wei et al. designed harmonic dilated convolution using the prior knowledge about the harmonic structure, which improved the model's performance [24]. However, each dilated convolution could only extract one kind of frequency relationship (fundamental frequency and ith overtone). They used 8 dilated convolutions but still couldn't model a complete harmonic structure. The attention mechanism is suitable for modeling the global dependencies of harmonic series. Wu et al. [25] used a masked self-attention to capture the harmonics up to 8 times, which is similar to the method in [24].

In the release phase of the note, the harmonics are less apparent, making the estimation of offset more challenging. In addition, the note's ending time is influenced by the piano pedals and playing techniques, so the annotations of some offsets are not time-aligned to the audio in existing datasets. These annotations limit the performance of supervised-learning networks.

Considering the aforementioned issues, the contributions of this paper are as follows:

1) A harmonic-aware frequency attention is proposed. The multi-head harmonic masks are designed based on the harmonic positions, which contain multiple possible harmonic series that include f_0. As shown in Fig. 1, Aggregating the residual non-harmonic components, the harmonic-aware attention captures the musical frequency structure of piano notes.
2) We propose a multi-task transcription model with the harmonic-aware frequency attention and local relative time attention. The time attention with learnable temporal attention range masks is applied to model the short-term temporal dependencies on different branches.
3) The dynamic behavior of modern piano pedals that affects on the note duration is considered in this study. A note duration modification method is proposed to locate the offset by tracking the state of the sustain pedal and sostenuto pedal.

Experiments on the MAESTRO dataset demonstrate that the proposed system achieves state-of-the-art transcription performance on both frame-wise and note-wise F1 metrics. The visualization of the attention matrices and results shows that the harmonic-aware frequency attention module can emphasize the harmonic structure of music, and the time attention models different temporal dependencies for the subtasks. By rectifying the frame and the offset labels with our note duration modification method, the transcription performance is further improved.

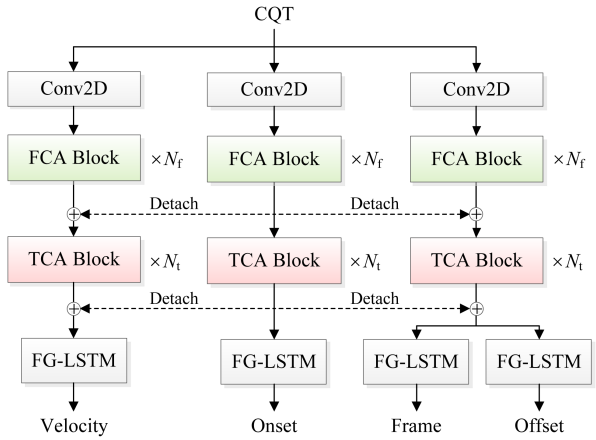

Fig. 2. The overall model architecture. $\oplus$ is element-wise summation operation.

The remainder of the paper is organized as follows. Section II presents the proposed piano transcription model, including the harmonic-aware frequency attention and the local relative time attention. Section III describes the note duration modification method. Section IV introduces the experimental settings. In Section V, the experimental results and discussions are provided. Finally, we conclude our work in Section VI.

II. Proposed Methods

A. Model Overview

The proposed harmonic-aware frequency attention and local relative time attention transcription network (HATANet) is designed as shown in Fig. 2. The piano transcription task in this study can be represented as:

$$\mathbf{Y}_q = \mathcal{F}(\mathbf{X}) \tag{1}$$

where the input $\mathbf{X} \in \mathbb{R}^{L\times F}$ is an L frame and F frequency bin CQT [26] spectrogram of the raw piano audio. Similar to the OAF system [27], the outputs $\mathbf{Y}_q \in \mathbb{R}^{L\times K}, q \in \{\text{frame,onset,offset,velocity}\}$ are the L frame and K piano pitch piano rolls of transcription subtasks. These subtasks include frame-level pitch estimation, onset detection, offset detection, and velocity estimation.

The losses for the frame, onset, and offset are calculated with binary cross-entropy. The loss for the velocity is a categorical cross-entropy loss. Specifically, the losses are:

$$l_{\text{onset}} = \sum_{k=1}^{88}\sum_{t=1}^{L} l_{\text{BCE}}(I_{\text{onset}}(t,k), P_{\text{onset}}(t,k)) \tag{2}$$

$$l_{\text{frame}} = \sum_{k=1}^{88}\sum_{t=1}^{L} l_{\text{BCE}}(I_{\text{frame}}(t,k), P_{\text{frame}}(t,k)) \tag{3}$$

$$l_{\text{offset}} = \sum_{k=1}^{88}\sum_{t=1}^{L} l_{\text{BCE}}(I_{\text{offset}}(t,k), P_{\text{offset}}(t,k)) \tag{4}$$

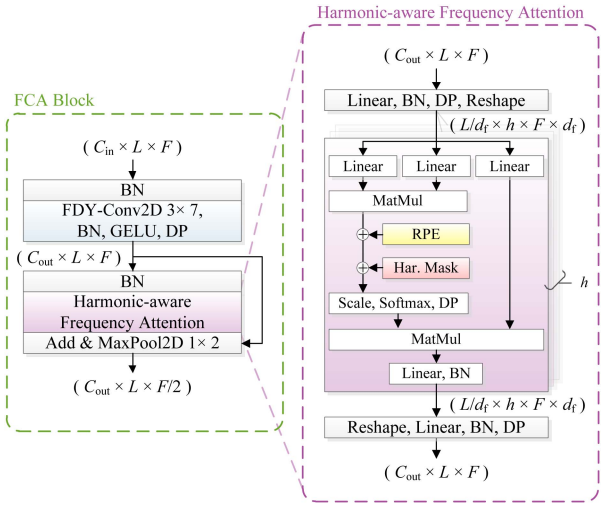

Fig. 3. The architecture of the frequency-convolution-attention (FCA) block. $\oplus$ is element-wise summation operation.

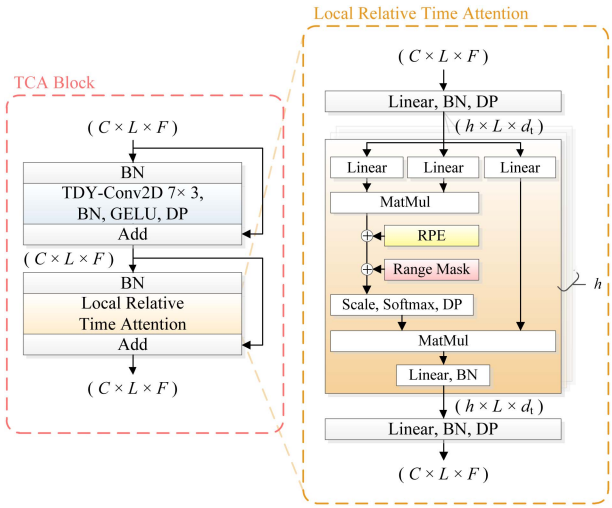

Fig. 4. The architecture of the time-convolution-attention (TCA) block. $\oplus$ is element-wise summation operation.

$$l_{\text{velocity}} = \sum_{k=1}^{128} \sum_{t=1}^{L} l_{\text{CCE}}(I_{\text{velocity}}(t,k), P_{\text{velocity}}(t,k)) \quad (5)$$

where l_{BCE} is the binary cross-entropy loss function, l_{CCE} is the categorical cross-entropy loss function, I and P are the ground-truth and model-predicted values, respectively. The total loss function is calculated by:

$$l = l_{\text{onset}} + l_{\text{frame}} + l_{\text{offset}} + l_{\text{velocity}} \quad (6)$$

The entire transcription system comprises the input convolution, the frequency-convolution-attention (FCA) blocks, the time-convolution-attention (TCA) blocks and the frequency grouped LSTMs (FG-LSTM) [24]. The FCA block is designed to capture the frequency structure of the harmonic and residual frequency components. As Fig. 3 shows, the FCA block contains a frequency dynamic convolution (FDY-Conv) layer [28] and a harmonic-aware frequency attention (HFA) layer. The TCA block is proposed to extract local time-variant attributes from neighboring audio frames. As shown in Fig. 4, the TCA block contains a temporal dynamic convolution (TDY-Conv) [29] layer and a local relative time attention (LRTA) layer. Each FG-LSTM has 88 frequency groups, which represent the 88 piano keys. All of the outputs are scaled by a sigmoid function, except for the velocity output.

B. Harmonic-Aware Frequency Attention

The attention mechanism [30] in neural networks learns the importance and correlation between each term in a sequence. In the FCA block, we use a self-attention mechanism to capture the frequency relationship of the musical acoustic structure. The harmonic-aware frequency attention (HFA) is applied to all of the F frequency bins for every d_f frame. On each attention head, the HFA can be represented as:

$$\text{Attention}(\mathbf{Q}_f, \mathbf{K}_f, \mathbf{V}_f) = \text{softmax}\left(\frac{\mathbf{Q}_f\mathbf{K}_f^\top + \mathbf{R}_f + \mathbf{M}_f}{\sqrt{d_f}}\right)\mathbf{V}_f \quad (7)$$

where $\mathbf{Q}_f, \mathbf{K}_f, \mathbf{V}_f \in \mathbb{R}^{F\times d_f}$ are the query, key and value matrices, $\mathbf{R}_f \in \mathbb{R}^{F\times F}$ is the global relative position embedding matrix and $\mathbf{M}_f \in \mathbb{R}^{F\times F}$ is the harmonic mask. The harmonic mask (HM) $\mathbf{M}_f$ is designed to extract the harmonic components.

The harmonic mask (HM) $\mathbf{M}_f$ is based on the position of the harmonics. The musical harmonic relationship between the fundamental frequency f_0 and kth harmonic f_m in the CQT spectrogram can be represented as:

$$f_m = kf_0 = 2^{\frac{m}{Q}} f_0, k = 1, 2, 3 \ldots \quad (8)$$

where Q denotes the number of bins in each octave. In the harmonic series, the distance m between the f_0 and f_m is:

$$m = Q \cdot \log_2 k \quad (9)$$

The harmonic series of f_0 when $Q = 12$ is shown in Fig. 5(a), where the indices on the axis are the bin intervals between the fundamental frequency and the overtones. The vertical lines on the positive semi-axis represent the harmonic components.

There are also other possible harmonic series that include f_0, such as the partials of fundamental frequency $\frac{1}{2}f_0$ and $\frac{1}{3}f_0$. To cover these possible harmonic series, an additional parameter anchor n is introduced and (8) is modified as follows:

$$f_{m'} = \frac{kf_0}{n} = 2^{\frac{m'}{Q}} f_0, k = 1, 2, 3 \ldots \quad (10)$$

The distance m' between the f_0 and kth harmonic $f_{m'}$ is:

$$m' = Q \cdot \log_2 \frac{k}{n} \quad (11)$$

The harmonic series of anchor $n = 2, 3, 4$ are shown in Fig. 5(b)–(d), which also include the harmonic components on the negative semi-axis. We can observe that some non-integer multiple frequency bins of f_0 are reserved in these series, such as $\frac{3}{2}f_0, (n = 2)$ and $\frac{2}{3}f_0, (n = 3)$.

The harmonic masks are set according to the distances between the harmonic components. The 2-dimension harmonic masks are added to the $\mathbf{Q}_f\mathbf{K}_f^\top$ attention matrix to retain the

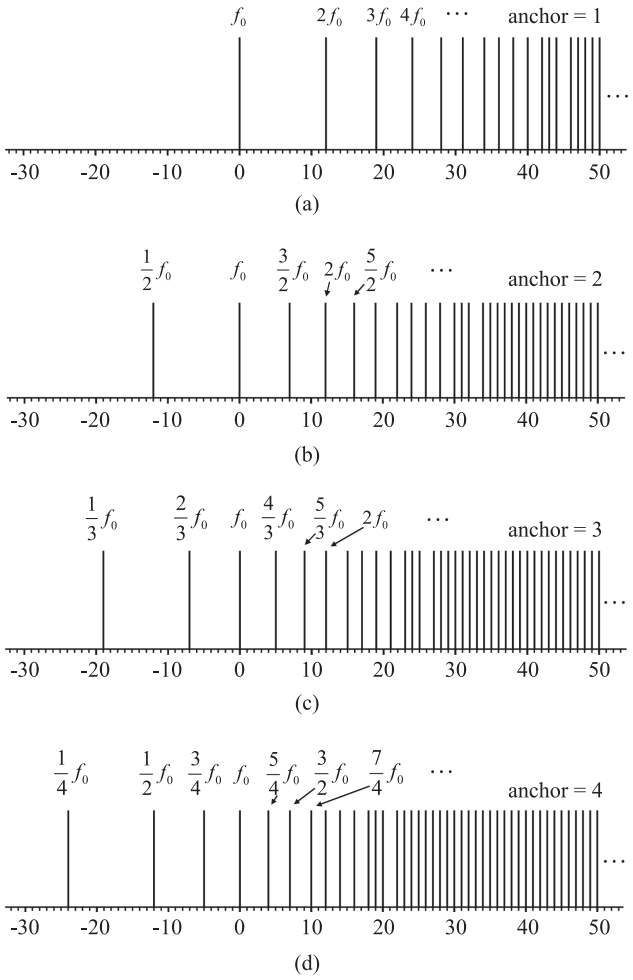

Fig. 5. The harmonic series of f_0 when $Q = 12$.

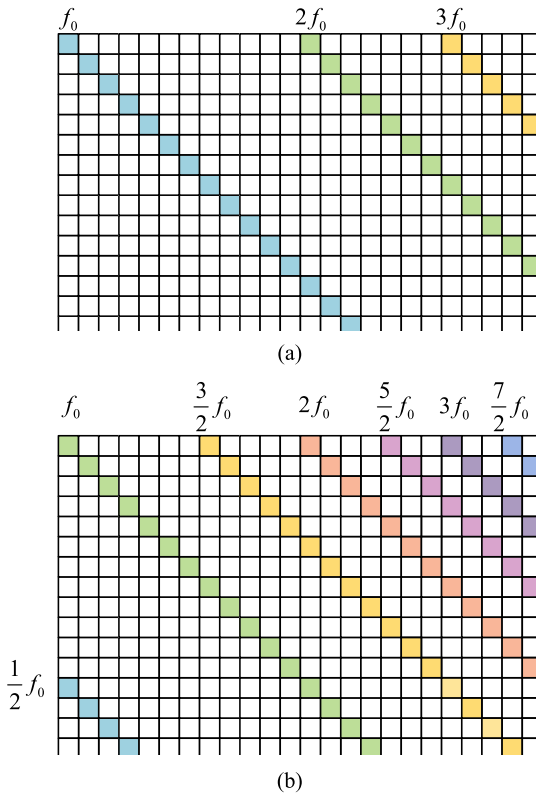

Fig. 6. The schematic diagram of the harmonic masks when $Q = 12$. (a) anchor $= 1$. (b) anchor $= 2$.

needed attention points. With $Q = 12$, Fig. 6 presents the schematic diagram of the harmonic masks when anchor $n = 1$ or 2. As shown in Fig. 6, only the painted points, where the key and query satisfy the harmonic distances in (11), are retained.

We also use the masks to extract the non-harmonic components, which correspond to the unpainted points in Fig. 6. All the harmonic masks $\mathbf{M}_{\mathrm{f}i}, i \in \{1, 2, \ldots, h\}$ for h attention heads can be divided into harmonic and residual non-harmonic groups. In the harmonic group, the HMs are constructed from different anchors and only the frequency bins in the harmonic series are retained, whereas the non-harmonic bins are discarded. In contrast, the HMs of the non-harmonic group retain only the non-harmonic frequency bins, while all the harmonic series bins are discarded. All the HMs in the harmonic and the residual non-harmonic groups separate the frequency domain into complementary sub-spaces.

The multi-head attention information is jointed by projecting the stack of each attention output, which can be represented as:

$$\begin{aligned}\mathrm{MultiHead}(\mathbf{Q}_{\mathrm{f}}, \mathbf{K}_{\mathrm{f}}, \mathbf{V}_{\mathrm{f}}) &= \mathrm{Stack}(\mathrm{head}_1; \ldots; \mathrm{head}_h)\mathbf{W}_{\mathrm{f}}^{O} \\ \text{where } \mathrm{head}_i &= \mathrm{Attention}(\mathbf{Q}_{\mathrm{f}}\mathbf{W}_{\mathrm{f}i}^{Q}, \mathbf{K}_{\mathrm{f}}\mathbf{W}_{\mathrm{f}i}^{K}, \mathbf{V}_{\mathrm{f}}\mathbf{W}_{\mathrm{f}i}^{V})\end{aligned} \tag{12}$$

where the projections are parameter matrices $\mathbf{W}_{\mathrm{f}i}^{Q} \in \mathbb{R}^{d_{\mathrm{f}} \times d_{\mathrm{f}}}$, $\mathbf{W}_{\mathrm{f}i}^{K} \in \mathbb{R}^{d_{\mathrm{f}} \times d_{\mathrm{f}}}$, $\mathbf{W}_{\mathrm{f}i}^{V} \in \mathbb{R}^{d_{\mathrm{f}} \times d_{\mathrm{f}}}$, and $\mathbf{W}_{\mathrm{f}}^{O} \in \mathbb{R}^{d_{\mathrm{f}} \times d_{\mathrm{f}}}$.

C. Local Relative Time Attention

In the TCA block, we use the self-attention mechanism to model the temporal relationship of adjacent audio frames. The local relative time attention (LRTA) is based on our previous work in [31], which can be seen as a "transposition" type of the HFA. The LRTA is applied to d_{t}-dimension adjacent local audio L frames. For each attention head, the LRTA can be represented as:

$$\mathrm{Attention}(\mathbf{Q}_{\mathrm{t}}, \mathbf{K}_{\mathrm{t}}, \mathbf{V}_{\mathrm{t}}) = \mathrm{softmax}\left(\frac{\mathbf{Q}_{\mathrm{t}}\mathbf{K}_{\mathrm{t}}^{\top} + \mathbf{R}_{\mathrm{t}} + \mathbf{M}_{\mathrm{t}}}{\sqrt{d_{\mathrm{t}}}}\right)\mathbf{V}_{\mathrm{t}} \tag{13}$$

where $\mathbf{Q}_{\mathrm{t}}, \mathbf{K}_{\mathrm{t}}, \mathbf{V}_{\mathrm{t}} \in \mathbb{R}^{L \times d_{\mathrm{t}}}$ are the query, key and value matrices, $\mathbf{R}_{\mathrm{t}} \in \mathbb{R}^{L \times L}$ is the local relative position embedding matrix and $\mathbf{M}_{\mathrm{t}} \in \mathbb{R}^{L \times L}$ is the attention range mask (ARM).

It has been demonstrated that the different temporal phases of the note evolution have distinctive importance for different transcription targets [13], [15], [27]. For example, the frame-level onset detection focuses more on short-term dependence than the offset detection. Therefore, the attention distance N in different subtasks on each head may be different. To obtain the proper N, we simplify the adaptive attention mask proposed in [32], and adopt it as the basis function $m_z(x)$ of the ARM in

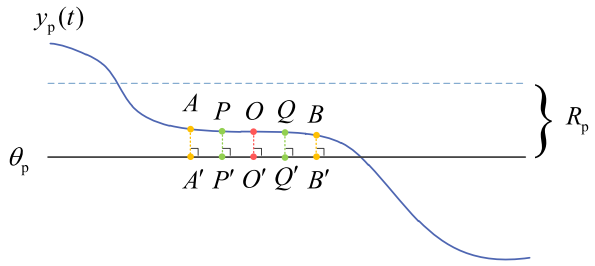

Fig. 7. The calculation demonstration diagram of the note duration modification method.

the bi-directional relative position embedding method:

$$m_z(x) = \min[\max(z - |x|, 0), 1], \quad x \in [-L+1, L-1], z \in [0, L-1] \tag{14}$$

where $m_z(x) \in [0, 1]$ and x refers to the directed relative distance in audio frame sequence L. z is the trainable parameter controlling the maximum attention relative distance N, where $N = 2z + 1$. To enable the attention mechanism to model multi-scale temporal dependence in different subtasks, $m_{zi}(x), i \in \{1, 2, \ldots, h\}$ are extended and formed as the ARMs $\mathbf{M}_{ti}, i \in \{1, 2, \ldots, h\}$ for each head in different branches.

Similar to (12), we use the same expressions but different notation t for multi-head time attention:

$$\begin{aligned} &\text{MultiHead}\left(\mathbf{Q}_t, \mathbf{K}_t, \mathbf{V}_t\right) = \text{Stack}\left(\text{head}_1; \ldots; \text{head}_h\right) \mathbf{W}_t^O \\ &\text{where head}_i = \text{Attention}\left(\mathbf{Q}_t \mathbf{W}_{ti}^Q, \mathbf{K}_t \mathbf{W}_{ti}^K, \mathbf{V}_t \mathbf{W}_{ti}^V\right) \end{aligned} \tag{15}$$

III. Note Duration Modification Method

In modern piano performances, piano pedals play an important role in polishing and enriching the tones that be played [33]. For example, the sustain pedal (i.e., damper pedal) is found to increase the decay time of notes [34]. Many existing transcription methods have considered the influence of sustain pedal on the piano notes. The pedal-extended note is judged to end when the pedal value is less than the pedal threshold. In existing transcription methods, the sustain pedal threshold is set to 64, approximately half the maximum MIDI control change value [13], [14], [16], [18], [24], [27], [35], [36].

In this study, we consider the various effects of the pedal techniques on the note duration. Apart from the sustain pedal, the sostenuto pedal also provides a selective sustaining effect, which allows the piano strings to continue vibrating freely after the keys are released. In addition, the playing techniques of pedals affect the duration and intensity of the sustained sound. We propose a note duration modification (NDM) method to locate the note ending time by tracking the state of the sustain pedal and the sostenuto pedal. The audio and the note labels can be aligned with a higher accuracy using the NDM-located ending time.

For the sustain pedal or the sostenuto pedal, the pedal speed value and the pedal acceleration value must be calculated to implement the NDM method. As shown in Fig. 7, $y_p(t)$ is the pedal value curve with respect to time t, θ_p is the pedal threshold, and R_p is the pedal detection range. The points $O(t_O, y_O)$, $A(t_1, y_1), B(t_2, y_2), P(t_3, y_3), Q(t_4, y_4)$ are points on the pedal value curve. The point O' is the middle point of $P'Q'$ as well as $A'B'$. The pedal speed at the point O is calculated as follows:

$$v_O = \frac{y_2 - y_1}{2\rho_1} = \frac{y_2 - y_1}{t_2 - t_1} \tag{16}$$

where $A'O' = B'O' = \rho_1$. Subsequently, the pedal speed value $v_p(t)$ on each time can be calculated. The pedal acceleration at the point O is defined as:

$$a_O = \frac{v_Q - v_P}{2\rho_2} = \frac{v_Q - v_P}{t_4 - t_3} \tag{17}$$

where $P'O' = Q'O' = \rho_2$. Then the pedal acceleration value $a_p(t)$ on each time can be calculated. For demonstration clarity, we define the beginning time of arbitrary piano note as t_{begin}, the ending time is t_{end}, and the note duration is $d_{\text{note}} = t_{\text{end}} - t_{\text{begin}}$, where $t_{\text{end}} > t_{\text{begin}}$. When the key is released without the pedal, t_{end} is the ending time of the key. When the pedal technique is involved in piano performance, the NDM rectifies the ending time of the note and the note duration in the following three situations.

1) The Notes Played With Slow Half-Pedal Technique: When the half-pedal is played, the player adjusts the pressure on the pedal to find a position where the dampers touch the strings but do not completely release them. To present soothing emotion in piano pieces, the half-pedal technique is involved to slowly weaken the piano strings vibration. For the notes played with slow half-pedal technique, the actual offset time of the note is earlier than when the pedal value reaches the pedal threshold. In this case, we end the long notes when the following conditions are satisfied:

$$\begin{cases} \theta_p \le y_p(t) \le \theta_p + R_p \\ -v_{\max} \le v_p(t) \le 0 \\ 0 \le a_p(t) \le a_{\max} \end{cases} \tag{18}$$

where $v_{\max}$ and $a_{\max}$ are the allowed maximum pedal speed value and pedal acceleration value, respectively.

2) The Notes Played With Quick Half-Pedal Technique: Quick pedal changes are often used for staccato passages or when a clear separation between notes is desired. When the note is played using the quick half-pedal technique, the transient contact of piano damper will end the note. Specifically, we end the notes if the following conditions are satisfied:

$$\begin{cases} \theta_p \le y_p(t) \le \theta_p + R_p \\ v_p(t) \le 0 \\ d_{\text{note}} \ge d_{\min} \end{cases} \tag{19}$$

where there is a minimum note duration $d_{\min}$ for avoiding fragmented short notes.

3) Pedal Value is Less Than the Pedal Threshold: For each pedal extended piano note, it will be finally ended when the pedal value is less than the pedal threshold:

$$y_p(t) < \theta_p \tag{20}$$

Algorithm 1: Note Duration Modification Based on The Sustain Pedal or The Sostenuto Pedal.

Input: $y_{\mathrm{p}}(t), v_{\mathrm{p}}(t), a_{\mathrm{p}}(t), \theta_{\mathrm{p}}, R_{\mathrm{p}}, v_{\max}, a_{\max}, d_{\min}$
Output: $t_{\mathrm{end}}|_{\{\text{ALL NOTES}\}}, d_{\mathrm{note}}|_{\{\text{ALL NOTES}\}}$

```
for NOTE ∈ {ALL NOTES} do
    d_note ← t_end − t_begin
    t ← t_end
    while y_p(t) ≥ θ_p do
        if reonset event then
            break
        else if y_p(t) ≤ θ_p + R_p and −v_max ≤ v_p(t) ≤ 0
        and 0 ≤ a_p(t) ≤ a_max then
            break
        else if y_p(t) ≤ θ_p + R_p and v_p(t) ≤ 0
        and d_note ≥ d_min then
            break
        end if
        t ← t + 1
    end while
    t_end ← t
    d_note ← t_end − t_begin
end for
```

This is the same as the previous pedal-involved note duration labeling strategy. This effective method is retained as the last one of the NDM methods.

Beyond the above three NDM methods, the key re-pressing event also ends the previous activated note. The complete calculation procedure is shown as in Algorithm 1.

IV. Experiments

A. Dataset

We use two well-known piano datasets in the experiments. The MAESTRO dataset [27] contains about 200 hours of high-quality realistic recordings and MIDI files. The MIDI data include key strike velocities and pedal positions. We use the provided train, validation, and test splits in the MAESTRO dataset. The train split of MAESTRO V3.0.0 is used for training and the test splits of different versions are used for evaluation.

The MAPS dataset [37] is also used to investigate the cross-domain robustness of the proposed model. The model trained on MAESTRO V3.0.0 is evaluated on the MAPS dataset. We use 60 audio recordings of realistic piano pieces in the "ENSTDkAm" and "ENSTDkCl" as the evaluation test split.

B. Preprocessing

We use the *librosa* [38] library for piano audio preprocessing. The raw audio is resampled to 16 kHz and clipped to 8-second pieces for each sample. The log amplitude CQT spectrogram is then calculated with a Hanning window, frame hop length of 20 milliseconds, minimum frequency of 27.5 Hz (A0), 48 bins per octave, total frequency bins number of 352. Finally, the dynamic range of the spectrogram is rescaled to [−40 dB, +40dB].

We use the *pretty_midi* [39] library for the MIDI file preprocessing. Each note is marked separately in four piano rolls, which include frame, onset, offset and velocity. We use the default preprocessing method and NDM method to obtain the offset time. In the default method, when the sustain pedal value is no less than 64, the note is extended until the same key is repressed or the sustain pedal is off. As for the preprocessing with the NDM method in this work, we use $\theta_{\mathrm{p}} = 64$, $R_{\mathrm{p}} = 1$, $v_{\max} = 0$, $a_{\max} = 0$, $A'O' = 2$ frames, $P'O' = 1$ frame, and $d_{\min} = 40$ frames.

C. Model Configurations

The kernel size of the input convolution layer is 7×7. For the dynamic convolutions, we follow the configurations in [28], [29], [40] to set the FDY-Conv layer and the TDY-Conv layer. There are 4 basis kernels in each dynamic convolution layer. During training, the softmax temperature in all the dynamic convolutions reduces from 31 to 1 linearly in the first 10 epochs. There are $N_{\mathrm{f}} = 2$ FCA blocks in each branch. In the HFA layer, there are the $h = 8$ attention heads to compute harmonic frequency attention on every $d_{\mathrm{f}} = 5$ audio frames. The computation method of $\mathbf{R}_{\mathrm{f}}$ in (7) on each attention head is same as the methods proposed in [41] to reduce the computation complexity. In the first FCA block, the channel size C_{in} of the input is 16, C_{out} is 32 and F is 352. In the Second FCA block, the output channel C_{out} is 64 and F is 176. The Q values for the two sets of HMs are respectively 48 and 24. We implement the harmonic masks with anchor $n = 1, 2, 3, \ldots, 7$ for 7 attention heads, and the last head with non-harmonic mask contains all the residual frequency bins are never shown in those harmonic groups.

There are total $N_{\mathrm{t}} = 4$ TCA blocks in each branch after the FCA blocks. For all the four TCA blocks, the channel size C is 64, and $F = d_{\mathrm{t}} = 88$ equals to the number of piano pitches. The computation method proposed in [41] is also used to reduce the computation complexity of $\mathbf{R}_{\mathrm{t}}$ in (13). There are also $h = 8$ attention heads in the LRTA layers. And the maximum attention relative distance is $N = 99$ frames. All of the trainable parameter $z_i, i \in \{1, 2, \ldots, h\}$ on each head is randomly initialized by a uniform distribution $U(0, \frac{N-1}{2})$.

The FG-LSTM is a one-layer bidirectional LSTM [42] with 64 units in each direction, and its configurations are the same as those proposed in [24].

D. Experiment Setup

Similar to the settings in [16], we use the high-resolution regressing onset and offset times for the onset and offset branches. The training is optimized using the Adam [43] optimizer with a mini-batch size of 4. We use a learning rate of 0.0005 for 250 k steps with early stopping. There are a total of 6.4 M trainable parameters of the HATANet. The model is trained on two GeForce RTX 3090 for about two weeks.

The evaluation is performed on every piano piece in the test split. The frame-level [44] and the note-level evaluation metrics in the *mir_eval* [45] library are used to obtain the precision, recall and F1 score. Following the tolerance range settings in the previous studies [13], [16], [24], we use ±50 ms as the onset tolerance. We also use an offset tolerance of ±50 ms or 20% of the total note duration, whichever is greater. The velocity estimation includes a tolerance within 0.1 of the prediction and

TABLE I
EVALUATION RESULTS ON THE MAESTRO DATASET

Model	Params	FRAME			NOTE			NOTE W/ OFFSET			NOTE W/ OFF. & VEL.		
		P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)
		MAESTRO V1.0.0											
OAF [27]	26M	92.11	88.41	90.15	98.27	92.61	95.32	82.95	78.24	80.50	79.89	75.37	77.54
GAN [14]	26M	93.10	89.80	91.40	98.10	93.20	95.60	83.50	79.30	81.30	82.30	78.20	80.20
Transformer [18]	54M	-	-	-	98.62	93.46	95.95	85.77	81.31	83.46	84.45	80.07	82.18
HATANet(ours)	6.4M	95.77	94.23	**94.99**	99.11	96.70	**97.89**	92.31	90.08	**91.18**	91.03	88.84	**89.92**
		MAESTRO V2.0.0											
HPT [16]	20M	88.71	90.73	89.62	98.17	95.35	96.72	83.68	81.32	82.47	82.10	79.80	80.92
HPPNet [24]	1.2M	92.36	93.46	92.86	98.31	96.18	97.21	85.36	83.54	84.41	83.85	82.08	82.93
Semi-CRFs [35]	9M	93.85	88.72	91.11	98.66	94.50	96.51	90.68	86.89	88.72	89.68	85.96	87.75
HATANet(ours)	6.4M	95.24	93.03	**94.12**	98.86	96.46	**97.65**	91.64	89.40	**90.51**	90.30	88.08	**89.18**
		MAESTRO V3.0.0											
Transformer [18]	54M	-	-	-	98.61	93.60	96.01	86.19	81.86	83.94	84.95	80.70	82.75
HPPNet [24]	1.2M	92.79	93.59	93.15	98.45	95.95	97.18	84.88	82.76	83.80	83.29	81.24	82.24
Semi-CRFs [35]	9M	93.79	88.36	90.75	98.69	93.96	96.11	90.79	86.46	88.42	89.78	85.51	87.44
hFT-Transformer [21]	5.5M	92.82	93.66	93.24	99.64	95.44	97.44	92.52	88.69	90.53	91.43	87.67	**89.48**
HATANet(ours)	6.4M	95.51	92.62	**94.04**	99.05	96.61	**97.81**	91.70	89.47	**90.57**	90.35	88.17	89.25

HATANet is trained on the MAESTRO V3.0.0 train split and evaluated on different versions test splits.

ground truth after mapping the velocity values to [0, 1]. We use the heuristic method from previous researches [13], [16], [24], [27] to decode the note events from the output post-probability piano rolls. Only the onset and the frame outputs are used in the inference, and the onset and the frame thresholds are 0.275 and 0.475, respectively. The thresholds are decided based on the result of the validation set, using a heuristic grid searching method. During the inference, if the top-1 argmax velocity prediction on onset valid point refers to 0, the second argmax velocity prediction is chosen as the final output.

V. RESULTS

A. Evaluation Results of Piano Transcription

Table I shows evaluation results on different versions of test splits. The results of this work are compared with those of the previous researches: the dual-objective onsets and frames (OAF) model [27]; the generative adversarial networks (GAN) improved OAF model [14]; the sequence-to-sequence generic Transformer transcription model [18]; the harmonic convolution and frequency grouped LSTM (HPPNet) model [24]; the semi-Markov conditional random fields (Semi-CRFs) based transcription model [35]; and the hierarchical frequency-time Transformer (hFT-Transformer) [21]. Among the previous researches, the hFT-Transformer achieved all the best frame and note F1 scores.

The proposed HATANet achieves state-of-the-art transcription F1 scores on both frame-wise and note-wise metrics. Even compared with the hFT-Transformer on MAESTRO V3.0.0, the HATANet respectively improves the frame F1 score of 0.8%, the onset F1 score of 0.37% and the note with offset F1 score of 0.03%, reaching 94.04% frame F1 score, 97.81% onset F1 score, 90.57% note onset with offset F1 score. In Fig. 8, we compare the transcription posteriorgram output of our model with those of the OAF, the HPT, the Transformer, the HPPNet, and the hFT-Transformer for the following input example. We can observe that the HATANet's frame-wise output is closer to the ground truth, especially the transcriptions around 2 seconds and 12 seconds. The HATANet gives more accurate and confident predictions at concurrent pitches and sustained long-term notes. Another example in Fig. 9 shows the onsets transcription output from a different audio clip. It demonstrates that the HATANet also performs better on concurrent legato notes than the baselines, such as the transcriptions at around 6 seconds and 12 seconds.

The cross-domain evaluation results on the MAPS dataset are shown in Table II. Similar to the cross-domain examination in the OAF [27] and HPPNet [24], the model is trained on the MAESTRO dataset without data augmentation and evaluated on the MAPS dataset. We also evaluate the hFT-Transformer for comparison. As shown in Table II, the proposed HATANet achieves the best note-level onset and note-level onset with offset F1 scores, which are separately 88.60% and 69.50%. The frame-level pitch estimation metric of the HATANet is similar to that of the HPPNet. The results indicate that the proposed HATANet has good generalization ability on piano transcription tasks except the velocity estimation.

To evaluate the performance of the HATANet on low-resource dataset, we train the HATANet with randomly selected 30% and 10% samples of the MAESTRO V3.0.0 train split. The model is evaluated using the complete MAESTRO V3.0.0 test split. The evaluation results are shown in Table III. The OAF and the HPPNet low-resource evaluation results are referred from the paper of the HPPNet [24]. When the training set size decreases, the HATANet drops less than the OAF and HPPNet on the frame F1 and note F1 scores. Even though only 10% of the data is used for training, the HATANet achieves higher scores on all the note-level with offset metrics and note-level with offset & velocity metrics than the OAF and HPPNet. This demonstrates that the HATANet is robust in the low-resource scenario.

B. Ablation Studies

We conduct the ablation experiments by orderly canceling the HMs (i.e., retaining the global frequency attention, FA), the harmonic-aware frequency attention and the local relative time attention of the HATANet. The ablation results

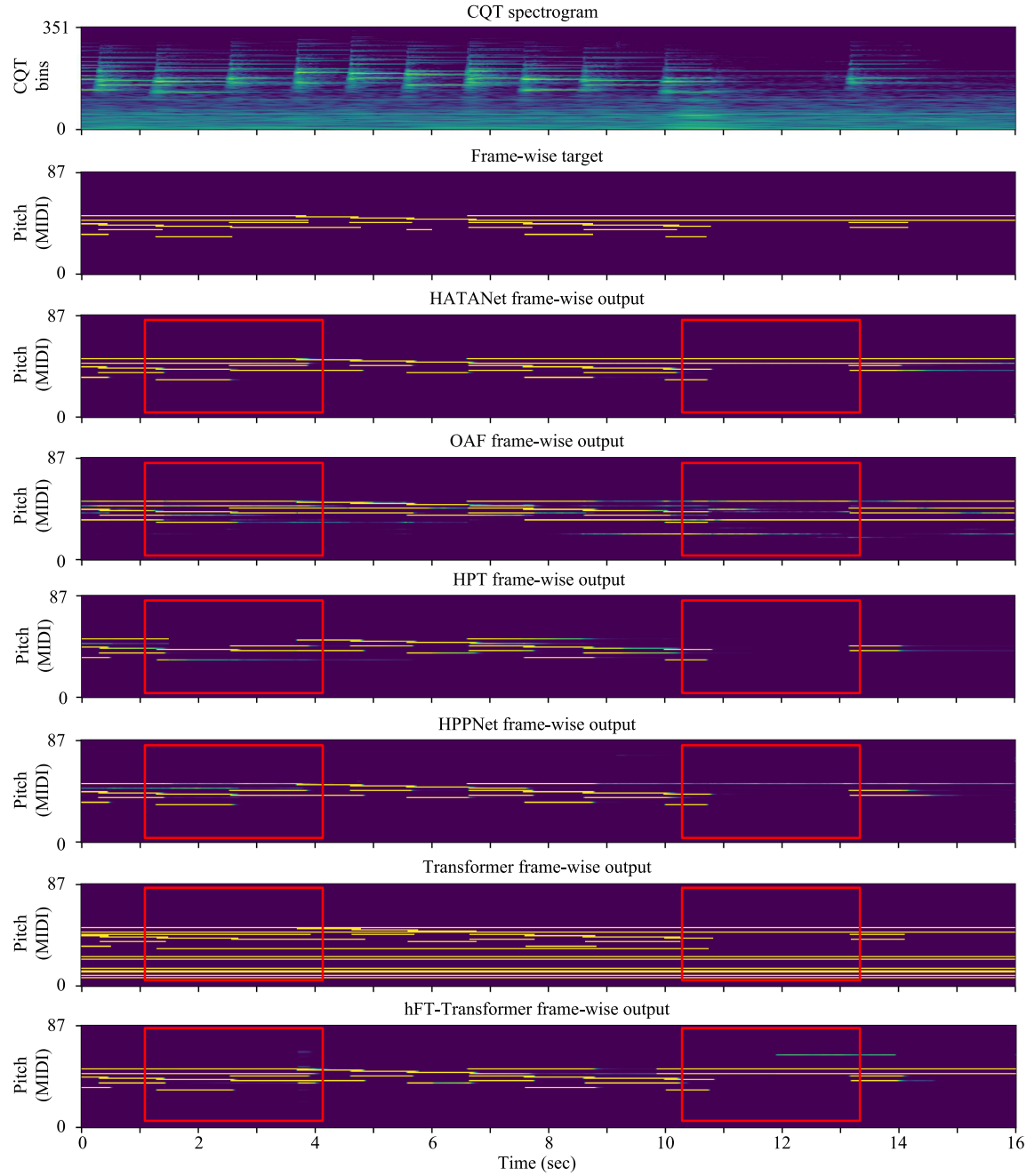

Fig. 8. Comparisons between the transcription output of the HATANet and the baselines. From the top to bottom: CQT spectrogram; ground truth frame-wise target; HATANet frame-wise output; OAF [27] frame-wise output; HPT [16] frame-wise output; HPPNet [24] frame-wise output; Transformer [18] frame-wise output; hFT-Transformer [21] frame-wise output. The frame-wise output of the Transformer is re-plotted from its MIDI-like output. The music segment is: MAESTRO V3.0.0/2018/MIDI-Unprocessed_Recital13-15_MID–AUDIO_15_R1_2018_wav-1.wav, 8'16.

TABLE II
CROSS-DOMAIN EVALUATION RESULTS ON THE MAPS DATASET

Model	FRAME			NOTE			NOTE W/ OFFSET			NOTE W/ OFF. & VEL.		
	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)
OAF [27]	-	-	80.02	-	-	83.04	-	-	61.84	-	-	48.07
HPPNet [24]	88.42	86.81	**87.56**	91.61	82.38	86.63	65.01	63.84	64.39	60.35	59.26	**59.77**
hFT-Transformer	77.69	72.00	74.74	82.78	85.42	82.52	57.80	60.10	57.95	38.90	40.22	38.74
HATANet(ours)	90.45	84.19	87.11	90.59	86.77	**88.60**	71.02	68.09	**69.50**	44.83	43.02	43.89

The models are trained on the MAESTRO V3.0.0 dataset and evaluated on the MAPS dataset without data augmentation.

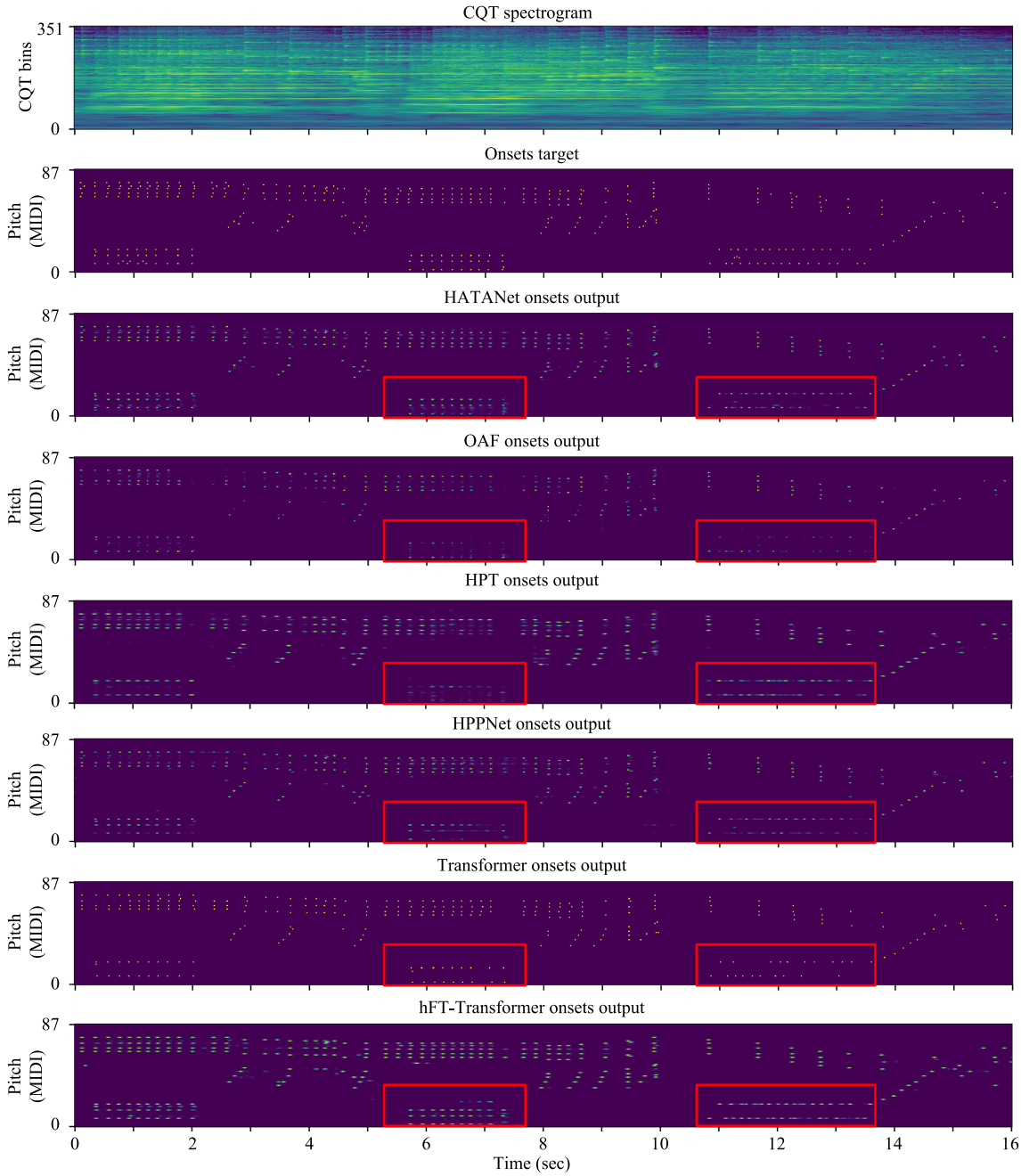

Fig. 9. Comparisons between the onsets output of the HATANet and the baselines. From the top to bottom: CQT spectrogram; ground truth onsets target; HATANet onsets output; OAF [27] onsets output; HPT [16] onsets output; HPPNet [24] onsets output; Transformer [18] onsets output; hFT-Transformer [21] onsets output. The music segment is: MAESTRO V3.0.0/2006/MIDI-Unprocessed_16_R1_2006_01-04_ORIG_MID–AUDIO_16_R1_2006_02_Track02_wav, 6'24.

listed in Table IV demonstrate the HATANet is superior to other ablated models and the HMs are effective. To evaluate the contribution of the harmonic components and non-harmonic components, we also train the HATANet with only the harmonic component masks of anchor $n = 1, 2, \ldots, 8$ applied on all the attention heads. This means that only harmonic series are retained in these masks. As shown in Table V, the model with full version of HMs outperforms the model with the "only harmonic" masks.

To further clarify the effectiveness of the harmonic mask and the harmonic frequency attention, we compare the attention matrices in each HFA head with and without the HMs. These attention matrices are averaged over a randomly chosen 64 s audio clip. As shown in Fig. 10, the HMs masked frequency attention captures the harmonic features of the overtones and the sub-harmonic frequencies. The relationship between the fundamental frequency and its harmonic series (anchor $n = 1$) is shown in Fig. 10(a). The subfigure indicates there are significant attention relationships between the f_0 and its harmonic series up to the 11th overtone. We can observe the "shifted" harmonic series (i.e., anchor $n = 2, 3, 4, \ldots, 7$) in Fig. 10(b)–(g). In these heads, the model is guided to treat each frequency bin as one

TABLE III
EVALUATION RESULTS ON THE MAESTRO V3.0.0 DATASET

Model	Data	FRAME			NOTE			NOTE W/ OFFSET			NOTE W/ OFF. & VEL.		
		P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)	P(%)	R (%)	F1(%)
OAF [24]	100%	94.43	85.50	89.68	98.67	92.08	95.22	82.25	76.88	79.44	80.80	75.55	78.05
	30%	91.14	80.00	85.08	98.13	89.57	93.60	-	-	-	-	-	-
	10%	90.66	72.70	80.36	96.64	85.46	90.62	-	-	-	-	-	-
HPPNet [24]	100%	92.79	93.59	93.15	98.45	95.95	97.18	84.88	82.76	83.80	83.29	81.24	82.24
	30%	90.72	92.11	91.35	96.46	94.46	95.43	-	-	-	-	-	-
	10%	92.10	84.96	88.31	96.59	90.94	93.52	-	-	-	-	-	-
HATANet(ours)	100%	95.51	92.62	94.04	99.05	96.61	97.81	91.70	89.47	90.57	90.35	88.17	89.25
	30%	92.91	91.87	92.32	98.61	95.40	96.96	88.73	85.88	87.27	85.87	83.14	84.46
	10%	92.51	89.94	91.14	97.57	94.79	96.13	86.62	84.20	85.37	83.69	81.37	82.49

The models are trained on 100%, 30%, and 10% of the MAESTRO v3.0.0 training split.

TABLE IV
ABLATION RESULTS ON THE MAESTRO V3.0.0 DATASET

Reserved Part			FRAME			NOTE			NOTE W/ OFFSET			NOTE W/ OFF. & VEL.		
HFA	FA	LRTA	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)
✓		✓	95.51	92.62	**94.04**	99.05	96.61	**97.81**	91.70	89.47	**90.57**	90.35	88.17	**89.25**
	✓	✓	94.23	92.49	93.31	99.55	95.33	97.36	90.83	87.03	88.87	88.06	84.41	86.17
✓			93.38	90.41	91.81	98.31	94.01	96.08	87.73	83.97	85.78	81.21	77.77	79.43
	✓		93.55	90.39	91.89	98.70	93.34	95.90	88.00	83.31	85.55	82.33	77.96	80.05
		✓	95.14	88.13	91.44	99.35	93.81	96.46	89.39	84.47	86.83	85.82	81.16	83.38
			91.12	78.54	84.25	93.95	84.72	89.01	78.17	70.58	74.11	72.23	65.23	68.49

TABLE V
COMPARISON OF THE HATANET WITH DIFFERENT HARMONIC MASKS

Model	FRAME			NOTE			NOTE W/ OFFSET			NOTE W/ OFF. & VEL.		
	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)
-Full version of HMs	95.51	92.62	**94.04**	99.05	96.61	**97.81**	91.70	89.47	**90.57**	90.35	88.17	**89.25**
-Without HMs	94.23	92.49	93.31	99.55	95.33	97.36	90.83	87.03	88.87	88.06	84.41	86.17
-Only harmonic components	95.34	91.64	93.40	99.41	95.64	97.47	90.80	88.06	89.39	88.68	86.00	87.31
-Only residual components	94.74	92.33	93.46	99.09	96.29	97.65	91.30	88.10	89.65	89.92	86.79	88.30

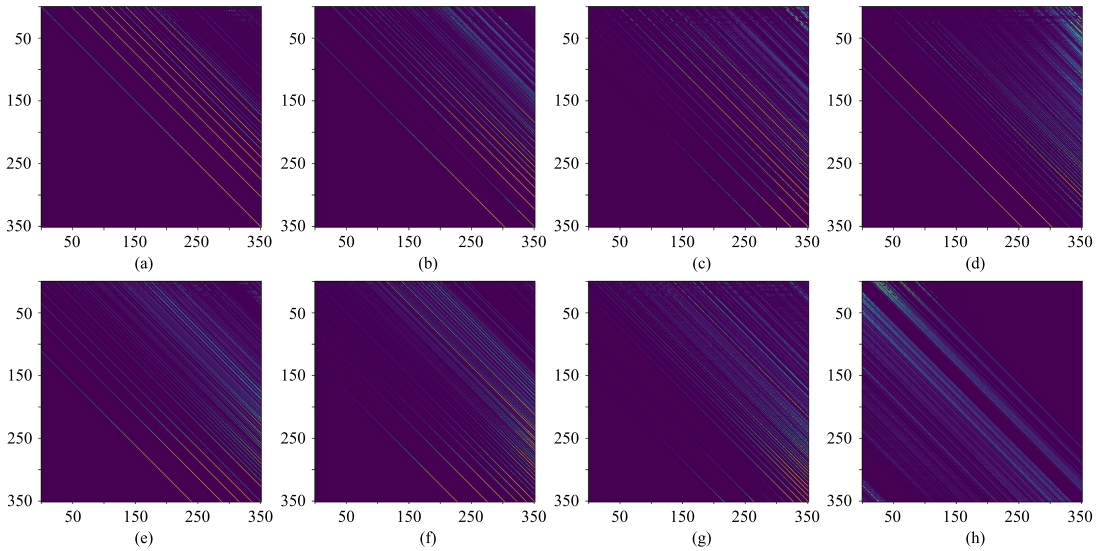

Fig. 10. The attention matrices in the HFA layer trained with the $Q = 48$ (in (10)) harmonic masks. The maximum frequency attention range is 351 frequency bins. The bright yellow points refer to high attention value. (a)–(g) are masked attention trained with the HMs in the harmonic group, where the anchor $n = 1, 2, 3, \ldots, 7$, respectively. (h) is masked attention trained with the corresponding HMs in the residual non-harmonic group.

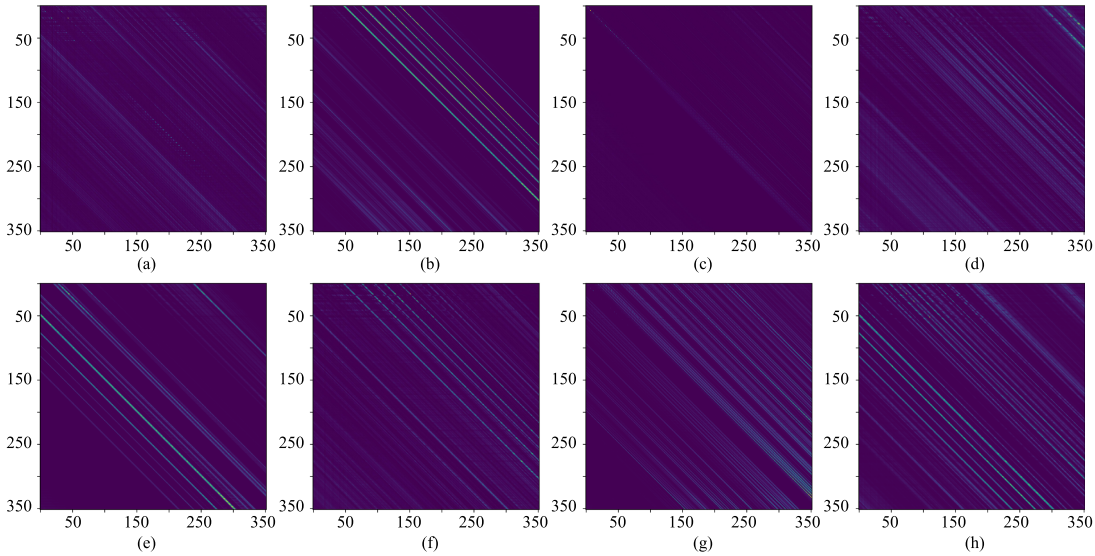

Fig. 11. The attention matrices in the HFA layer trained without harmonic masks. The maximum frequency attention range is 351 frequency bins. The bright yellow points refer to high attention value. (a)–(h) are attention matrices of eight heads, respectively.

of the overtones instead of the fundamental frequency. For example, there is a strong attention relationship between the 4th harmonic and the 2nd harmonic in Fig. 10(d). The model considers more on lower harmonic partials in higher frequency bins of larger anchors (Fig. 10(e) and (f)). As shown in Fig. 10(h), the residual non-harmonic components also have considerable effects in frequency attention. Apart from some useful non-harmonic components for pitch estimation, these residuals might include other possible harmonic series with different anchor. Moreover, these harmonic-beyond significance may be due to the realistic piano property on frequency deviation, which can be learned by the network for pitch modeling.

As shown in Fig. 11, the attention heads trained without the HMs also exhibit the ability to capture the structured frequency correlations. In Fig. 11(b), (e), and (h), we can observe some harmonic-like frequency structures naturally learned by the model, even without the guidance of masks. However, in other attention heads, these attention distributions are mostly blurrier than those trained with the HMs and lack a clear discrimination of harmonic features. An overview comparison between the attention trained with and without the HMs is shown in Fig. 12. Although they both model the multiple relationships among all the frequency bins, the HFA with HMs shows a more confident attention pattern on the musical harmonic structure. We use a multipitch example of C4 and G4 to detail the differences between the attention patterns described above. The spectrums passing through the attention layer with and without the HMs are shown in Fig. 13. The attention output with the HMs presents higher amplitudes on the harmonic components and depresses other non-harmonic components. This emphasized attention pattern with the HMs facilitate the model to capture notes from structured musical frequencies more effectively.

We also compare the average attention values in the LRTA heads of the different branches. These attention values are averaged on a randomly chosen 64 s audio clip over all the attention heads across all the LRTA layers. As shown in Fig. 14, compared

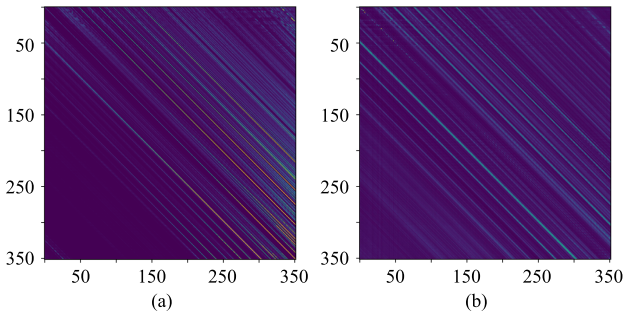

Fig. 12. The average of the attention matrices across all the attention heads trained with and without the HMs. The maximum frequency attention range is 351 frequency bins. The bright yellow points refer to high attention value. (a) is the average of the attention heads in Fig. 10. (b) is the average of the attention heads in Fig. 11.

to the more compact attention range of the onset branch, the network considers more distant temporal dependencies in the frame & offset branch. Considering the evolution of a piano note, these different attention distance preferences are reasonable. For example, the energy increases rapidly at the onset, and each frame is more relevant to neighboring frames in the audio sequence.

C. Evaluation With the Note Duration Modification Method

The NDM method is evaluated on the MAESTRO V3.0.0 dataset. The new annotations of notes are obtained using the NDM method. When the NDM-obtained duration differs from its original duration (obtained using the default method [13]) beyond a certain tolerance, we term this note "changed". The numbers of changed notes are listed in Table VI. There are 5 types of tolerances: 5 ms, 10 ms, 20 ms, 50 ms, and the mir_eval default tolerance (i.e., 50 ms or 20% note duration whichever is greater). In the original MAESTRO dataset, the audio and MIDI files are aligned with an accuracy of approximately 3 ms.

TABLE VI
THE CHANGED NOTES BETWEEN THE DEFAULT DATASET AND THE NDM DATASET

	Notes num.	Changed notes number and percentage under offset tolerances.									
		5ms		10ms		20ms		50ms		50ms & 20%	
Train split	5659329	231322	4.09%	229152	4.05%	108467	1.92%	76060	1.34%	28630	0.51%
Test split	741410	29203	3.94%	28884	3.90%	14205	1.92%	9659	1.30%	3254	0.44%
Validation split	639425	25940	4.06%	25675	4.02%	10219	1.60%	6831	1.07%	2602	0.41%
Total	7040164	286465	4.07%	283711	4.03%	132891	1.89%	92550	1.31%	34486	0.49%

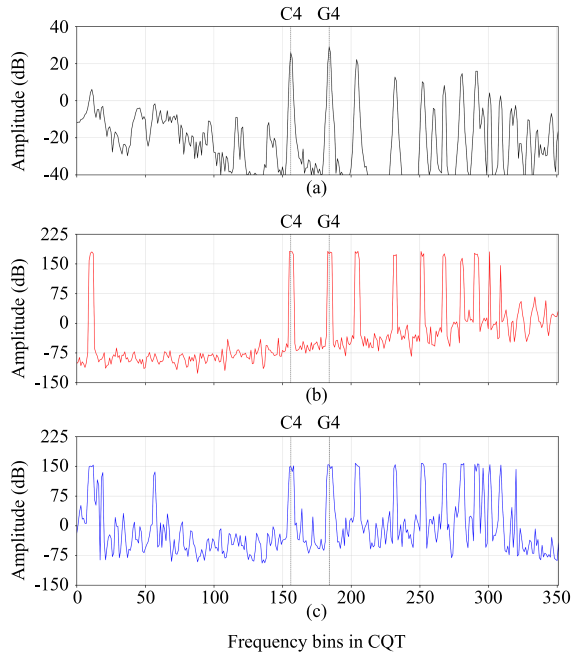

Fig. 13. The amplitude of the input and output on the first frequency attention layer with and without the HMs. The fundamental frequency bins of C4 and G4 are marked with dash line above. (a) The input spectrum, (b) the attention output with the HMs, (c) the attention output without the HMs.

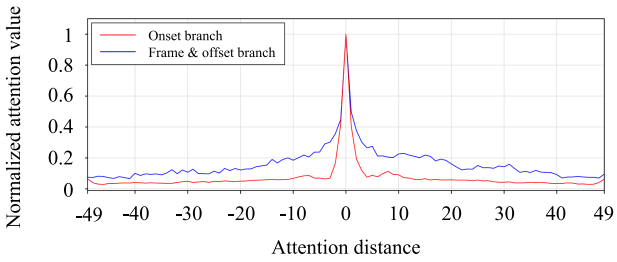

Fig. 14. The average attention values of the attention matrices over all the attention heads across all the LRTA layers.

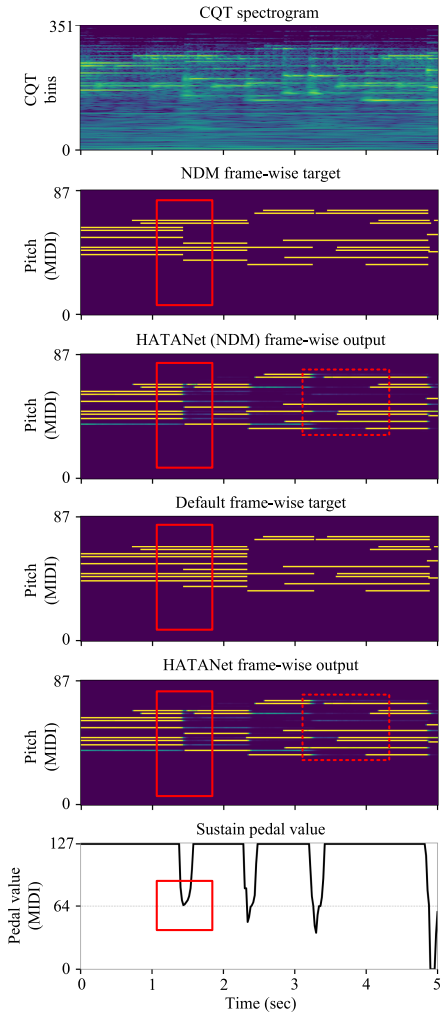

Fig. 15. Comparisons between the default frame-wise target and the NDM frame-wise target. From the top to bottom: CQT spectrogram; NDM frame-wise target; HATANet (trained with the NDM dataset) frame-wise output; default frame-wise output; HATANet (trained with the default dataset) frame-wise output; sustain pedal value changing with time. The music segment: MAESTRO V3.0.0/2014/MIDI-UNPROCESSED_09-10_R1_2014_MID–AUDIO_09_R1_2014_wav–4, 2'34.

With a time tolerance of 5 ms, 286465 notes have changed their annotations of offset by the NDM method, reaching about 4%.

Table VII presents the transcription results using different annotations. Due to the onset targets are not changed in the NDM dataset, the onset branch is frozen during the training, and its evaluation results are omitted in this section. As shown in the third row of Table VII, the pedal-rectified note ending targets further improved the HATANet performance on the frame F1 score of 0.3%, the onset with offset F1 score of 0.3% and the onset with offset & velocity F1 score of 0.32%, reaching 94.34%, 90.87% and 89.57%, respectively. Considering that 0.49% notes are changed under the *mir_eval* tolerance in Table VI, the transcription F1 score improvements brought by the NDM method are acceptable. As shown in the second row, when the HATANet is trained using the default dataset and evaluated using the NDM dataset, the results also exhibit performance improvements. These bonuses indicate that the

TABLE VII
EVALUATION RESULTS ON THE MAESTRO V3.0.0 DATASET WITH THE NOTE DURATION MODIFICATION METHOD

Model	Data	FRAME			NOTE W/ OFFSET			NOTE W/ OFF. & VEL.		
		P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)
HATANet	-Train & test on default	95.51	92.62	94.04	91.70	89.47	90.57	90.35	88.17	89.25
	-Train on default, test on NDM	95.42	93.00	94.19	91.87	89.64	90.74	90.53	88.34	89.42
	-Train & test on NDM	95.36	93.35	**94.34**	92.08	89.70	**90.87**	90.76	88.42	**89.57**
hFT-Transformer [21]	-Train & test on default	92.82	93.66	93.24	92.52	88.69	90.53	91.43	87.67	89.48
	-Train on default, test on NDM	91.40	94.21	92.79	92.76	88.96	90.79	91.62	87.90	**89.69**
	-Train & test on NDM	95.37	92.44	**93.82**	92.30	89.45	**90.83**	91.03	88.25	89.60

TABLE VIII
PEDAL TRANSCRIPTION EVALUATED RESULTS ON THE MAESTRO DATASET

Model	Params	FRAME			EVENT			EVENT W/ OFFSET		
		P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)
Liang (Implemented in [16])	13K	74.29	90.01	79.12	-	-	-	-	-	-
OAF (Implemented in [16])	20M	94.30	94.42	**94.25**	93.20	90.26	91.57	86.94	84.28	85.47
HPT [16]	14M	94.30	94.42	**94.25**	91.59	92.41	91.86	86.36	87.02	86.58
HATANet(ours)	2.2M	93.76	92.39	92.97	92.87	91.16	**91.90**	88.37	86.80	**87.48**

network naturally learns the frame-wise pitch and offset from a large amount of data, which are closer to the NDM targets. We also evaluate the hFT-Transformer [21] using the NDM test data, and the results also show the similar tendencies on the offset-related metrics. After training with the NDM dataset, the hFT-Transformer gains performance improvement on the frame-level, the note-level onset with offset, and the note-level onset with offset & velocity transcription results. Although the note-level onset with offset & velocity transcription F1 score of the NDM-trained hFT-Transformer is not the best, the recall score is higher than the default dataset-trained result.

Fig. 15 shows an audio piece played with a fast half-pedal technique. The second row of Fig. 15 shows the new frame-wise target modified by the NDM method. The third row shows the output of the HATANet, which is trained using the NDM dataset. The default target and output are shown in the forth and fifth rows, respectively. As shown in the last row of Fig. 15, the temporary variation of the sustain pedal value around 1.5 seconds indicates that there is a quick contact of the piano damper. It mutes all the pedal sustained notes at the same time and causes a subtle interruption of notes activations on the spectrogram. The original binary threshold labeling method omits this dynamic behavior of the piano pedal and causes a mismatch to the realistic spectrogram. In the NDM method, the interruption is considered. Moreover, the lower output of the HATANet (NDM), around 3.5 seconds, indicates that the HATANet trained with the NDM dataset performs more accurately on frame-wise transcription.

D. Sustain Pedal Transcription

We conduct a sustain pedal transcription experiment to evaluate the proposed attention-based method. The experiment of sustain pedal transcription are similar to the HPT pedal transcription in [16]. Based on the proposed HATANet, the sustain pedal transcription system is similar to the piano note transcription system. The difference is that the sustain pedal transcription system has only one branch, and the output shape is $T \times 1$ instead of $T \times 88$. For the sustain pedal frame-wise targets, the sustain pedal integer MIDI values from 0 to 127 are binarized according to 64, which MIDI values no less than 64 are regarded as "on" and the others are regarded as "off". The sustain pedal event-wise targets, loss functions and evaluation metrics are the same as those in the HPT pedal transcription system.

The evaluation results of pedal transcription are shown in Table VIII. The proposed HATANet-based pedal transcription model achieves the best sustain pedal event-wise onset and event-wise onset with offset F1 scores, which are separately 91.90% and 87.48%. The frame-wise sustain pedal transcription F1 score is lower than the OAF and the HPT, which is 93.66%. There are only 2.2 M parameters in the proposed model. Comparing to the baselines, the HATANet-based pedal transcription model has competitive transcription performance.

VI. CONCLUSION

We propose a multi-task transcription model with the harmonic-aware frequency attention and local relative time attention. The harmonic-aware frequency attention is designed to capture the musical frequency structure, especially the harmonic relationship. Many MIR tasks involving pitched instruments can also benefit from the frequency invariance on overtone characterizes. In the multi-task framework, the time attention with learnable temporal attention range masks can help the network model the temporal dependencies on different branches. We demonstrate that the proposed system achieves state-of-the-art transcription performance on both frame-wise and note-wise F1 metrics. In addition, we consider the influence of the piano pedals' dynamic behavior on note duration and propose a note duration modification method. The experimental results indicate that the modified transcription labels can further improve transcription performance.

There is scope for improvement in the future. The existing transcription methods aim to identify note's acoustic attributes, which are different from the actual playing processes. These actual performing behaviors, including the key and pedal playing,

are more challenging to recognize from raw audio and need to be studied further.

APPENDIX

The detailed model complexities of the proposed HATANet and OAF are shown in Table IX and Table X. A random input with a shape of [400 (frames), 352 (frequency bins)] is passed to the network to calculate the multiply-accumulate (MAC) operations. The OAF takes the Mel-scaled spectrograms as input, so its input size is [400 (frames), 229 (frequency bins)]. The counting process is implemented with the *ptflops* library. Notice that these complexity statistics only include the main operation layers in the onset branch.

TABLE IX
THE NUMBER OF MODEL PARAMETERS AND COMPLEXITY OF THE HATANET

Layer	Params	MACs
Input Conv2D	832	119.4 M
FDY-Conv2D	43.9 K	6.13 G
HFA	832	127.28 M
FDY-Conv2D	174.72 K	12.19 G
HFA	1.5 K	110.95 M
TDY-Conv2D	352.24 K	12.15 G
LRTA	32.71 K	148.97 M
TDY-Conv2D	352.24 K	12.15 G
LRTA	32.71 K	148.97 M
TDY-Conv2D	352.24 K	12.15 G
LRTA	32.71 K	148.97 M
TDY-Conv2D	352.24 K	12.15 G
LRTA	32.71 K	148.97 M
FG-LSTM	34.43 K	1.25 G
Onset Branch Total	1.79 M	69.15 G
Model Total	6.44 M	219.5 G

TABLE X
THE NUMBER OF MODEL PARAMETERS AND COMPLEXITY OF THE OAF

Layer	Params	MACs
Conv2D	576	57.16 M
Conv2D	20.88 K	1.92 G
Conv2D	41.76 K	1.92 G
Linear	4.2 M	1.68 G
BiLSTM	3.55 M	1.42 G
Linear	67.67 K	27.07 M
Onset Branch Total	7.88 M	7.02 G
Model Total	26.49 M	26.16 G

REFERENCES

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

[2] E. Benetos, S. Dixon, D. Giannoulis, H. Kirchhoff, and A. Klapuri, "Automatic music transcription: Challenges and future directions," *J. Intell. Inf. Syst.*, vol. 41, pp. 407–434, 2013.

[3] G. E. Poliner and D. P. W. Ellis, "A discriminative model for polyphonic piano transcription," *EURASIP J. Adv. Sign. Process.*, vol. 2006, pp. 1–9, 2006.

[4] G. Costantini, M. Todisco, R. Perfetti, R. Basili, and D. Casali, "SVM based transcription system with short-term memory oriented to polyphonic piano music," in *Proc. IEEE 15th Mediterranean Electrotechnical Conf.*, 2010, pp. 196–201.

[5] E. Benetos and T. Weyde, "An efficient temporally-constrained probabilistic model for multiple-instrument music transcription," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, 2015, pp. 701–707.

[6] A. Dessein, A. Cont, and G. Lemaitre, "Real-time polyphonic music transcription with non-negative matrix factorization and beta-divergence," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, 2010, pp. 489–494.

[7] N. Bertin, R. Badeau, and E. Vincent, "Enforcing harmonicity and smoothness in Bayesian non-negative matrix factorization applied to polyphonic music transcription," *IEEE Trans. Audio Speech Lang. Process.*, vol. 18, no. 3, pp. 538–549, Mar. 2010.

[8] T. F. Tavares, J. G. A. Barbedo, and R. d. F. Attux, "Unsupervised note activity detection in NMF-based automatic transcription of piano music," *J. New Music Res.*, vol. 45, pp. 118–123, 2016.

[9] A. Rizzi, M. Antonelli, and M. Luzi, "Instrument learning and sparse NMD for automatic polyphonic music transcription," *IEEE Trans. Multimedia*, vol. 19, pp. 1405–1415, 2017.

[10] N. Boulanger-Lewandowski, Y. Bengio, and P. Vincent, "Modeling temporal dependencies in high-dimensional sequences: Application to polyphonic music generation and transcription," in *Proc. Int. Conf. Mach. Learn.*, 2012, pp. 1159–1166.

[11] S. Böck and M. Schedl, "Polyphonic piano note transcription with recurrent neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2012, pp. 121–124.

[12] S. Sigtia, E. Benetos, and S. Dixon, "An end-to-end neural network for polyphonic piano music transcription," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 24, no. 5, pp. 927–939, May 2016.

[13] C. Hawthorne et al., "Onsets and frames: Dual-objective piano transcription," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, 2018, pp. 50–57.

[14] J. W. Kim and J. P. Bello, "Adversarial learning for improved onsets and frames music transcription," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, 2019, pp. 670–677.

[15] R. Kelz, S. Bock, and G. Widmer, "Deep polyphonic ADSR piano note transcription," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2019, pp. 246–250.

[16] 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 Trans. Audio Speech Lang. Process.*, vol. 29, pp. 3707–3717, 2021.

[17] K. W. Cheuk, Y.-J. Luo, E. Benetos, and D. Herremans, "Revisiting the onsets and frames model with additive attention," in *Proc. Int. Joint Conf. Neural Netw.*, 2021, pp. 1–8.

[18] C. Hawthorne, I. Simon, R. Swavely, E. Manilow, and J. Engel, "Sequence-to-Sequence piano transcription with transformers," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, 2021, pp. 246–253.

[19] J. P. Gardner, I. Simon, E. Manilow, C. Hawthorne, and J. Engel, "MT3: Multi-task multitrack music transcription," in *Proc. Int. Conf. Learn. Representation*, 2022, pp. 1–21.

[20] L. Ou, Z. Guo, E. Benetos, J. Han, and Y. Wang, "Exploring transformer's potential on automatic piano transcription," in *Proc. Int. Conf. Acoust. Speech Signal Process.*, 2022, pp. 776–780.

[21] K. Toyama, T. Akama, Y. Ikemiya, Y. Takida, W.-H. Liao, and Y. Mitsufuji, "Automatic piano transcription with hierarchical frequency-time transformer," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, 2023, pp. 1–8.

[22] W.-T. Lu, J.-C. Wang, M. Won, K. Choi, and X. Song, "SpecTNT: A time-frequency transformer for music audio," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, 2021, pp. 396–403.

[23] W. T. Lu, J.-C. Wang, and Y. N. Hung, "Multitrack music transcription with a time-frequency perceiver," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2023, pp. 1–5.

[24] W. Wei, P. Li, Y. Yu, and W. Li, "HPPNet: Modeling the harmonic structure and pitch invariance in piano transcription," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, 2022, pp. 709–716.

[25] R. Wu, X. Wang, Y. Li, W. Xu, and W. Cheng, "Piano transcription with harmonic attention," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2024, pp. 1256–1260.

[26] J. C. Brown, "Calculation of a constant Q spectral transform," *J. Acoustical Soc. Amer.*, vol. 89, no. 1, pp. 425–434, 1991.

[27] C. Hawthorne et al., "Enabling factorized piano music modeling and generation with the Maestro dataset," in *Proc. Int. Conf. Learn. Representation*, 2019, pp. 1–12.

[28] H. Nam, S.-H. Kim, B.-Y. Ko, and Y.-H. Park, "Frequency dynamic convolution: Frequency-adaptive pattern recognition for sound event detection," in *Proc. INTERSPEECH*, 2022, pp. 2763–2767.

[29] S.-H. Kim, H. Nam, and Y.-H. Park, "Temporal dynamic convolutional neural network for text-independent speaker verification and phonemic analysis," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2022, pp. 6742–6746.
[30] A. Vaswani et al., "Attention is all you need," in *Proc. 31st Int. Conf. Neural Inf. Proces. Syst.*, 2017, pp. 6000–6010.
[31] Q. Wang, M. Liu, X. Chen, and M. Xiong, "Multi-task piano transcription with local relative time attention," in *Proc. Asia Pacific Signal Inf. Process. Assoc. Annu. Summit Conf.*, 2023, pp. 966–971.
[32] S. Sukhbaatar, E. Grave, P. Bojanowski, and A. Joulin, "Adaptive attention span in transformers," in *Proc. Annu. Meeting Assoc. Comput. Linguistics*., 2020, pp. 331–335.
[33] S. P. Rosenblum, "Pedaling the piano: A brief survey from the eighteenth century to the present," *Perform. Pract. Rev.*, vol. 6, 1993, Art. no. 8.
[34] H.-M. Lehtonen, H. Penttinen, J. Rauhala, and V. Valimaki, "Analysis and modeling of piano sustain-pedal effects," *J. Acoustical Soc. Amer.*, vol. 122, no. 3, pp. 1787–1797, 2007.
[35] Y. Yan, F. Cwitkowitz, and Z. Duan, "Skipping the frame-level: Event-based piano transcription with neural semi-CRFs," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, vol. 25, pp. 20583–20595.
[36] T. Kwon, D. Jeong, and J. Nam, "Polyphonic piano transcription using autoregressive multi-state note model," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, 2020, pp. 454–461.
[37] V. Emiya, R. Badeau, and B. David, "Multipitch estimation of piano sounds using a new probabilistic spectral smoothness principle," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 18, no. 6, pp. 1643–1654, Aug. 2010.
[38] B. McFee et al., "Librosa: Audio and music signal analysis in Python," in *Pro. 14th Python Sci. Conf.*, 2015, pp. 18–24.
[39] C. Raffel and P. W. D. Ellis, "Intuitive analysis, creation and manipulation of MIDI data with pretty_midi," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, 2014, pp. 84–93.
[40] Y. Chen, X. Dai, M. Liu, D. Chen, L. Yuan, and Z. Liu, "Dynamic convolution: Attention over convolution kernels," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 11027–11036.
[41] Z. Dai, Z. Yang, Y. Yang, J. Carbonell, Q. V. Le, and R. Salakhutdinov, "Transformer-XL: Attentive language models beyond a fixed-length context," in *Proc. Annu. Meeting Assoc. Comput. Linguistics*., 2020, pp. 2978–2988.
[42] H. Sak, A. Senior, and F. Beaufays, "Long short-term memory recurrent neural network architectures for large scale acoustic modeling," in *Proc. INTERSPEECH*, 2014, pp. 338–342.
[43] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Representation*, 2015, pp 1–15.
[44] M. Bay, A. F. Ehmann, and J. Stephen Downie, "Evaluation of multiple-F0 estimation and tracking systems," in *Proc. Int. Soc. Music Inf. Retrieval Conf*., 2009, pp. 315–320.
[45] C. Raffel et al., "MIR_EVAL: A transparent implementation of common MIR metrics," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*s, 2014, pp. 367–372.

Qi Wang (Member, IEEE) received the B.S. and M.S. degrees from Shandong University, Jinan, China, in 2012 and 2015, respectively, and the Ph.D. degree in information and signal processing from the Institute of Acoustics, Chinese Academy of Sciences, Beijing, China, in 2018. She is currently a Lecturer with the Faculty of Information Technology, Beijing University of Technology, Beijing. Her research interests include audio signal processing, music signal processing, music information retrieval, and deep learning.

Mingkuan Liu received the B.E. degree in communication engineering from the Beijing University of Technology, Beijing, China, in 2021. He is currently working toward the M.S. degree in information and communication engineering with the Faculty of Information Technology, Beijing University of Technology. His research interests include audio signal processing, specifically applying deep learning, and signal processing techniques in the field of music information retrieval.

Changchun Bao (Senior Member, IEEE) is currently a Full Professor with the Faculty of Information Technology, Beijing University of Technology, Beijing, China. His research interests include speech and audio signal processing, speech coding, speech enhancement, speech transcoding, audio coding, audio enhancement, bandwidth extending for speech and audio signals, and 3D audio signal processing.

Maoshen Jia (Senior Member, IEEE) received the B.E. degree in electronic information engineering from Hebei University, Baoding, China, in 2005, the Ph.D. degree in electronic science and technology from the Beijing University of Technology, Beijing, China, in 2010. He is currently a Professor with the Faculty of Information Technology, Beijing University of Technology. His research interests include multichannel audio signal processing, audio coding, and array signal processing.