

Article

SpectTrans: Joint Spectral–Temporal Modeling for Polyphonic Piano Transcription via Spectral Gating Networks

Rui Cao ¹, Yan Liang ², Lei Feng ^{3,*}  and Yuanzi Li ³ ¹ College of Humanities, Xiamen Huaxia University, Xiamen 361000, China; caor@hxxu.edu.cn² College of Music, The Catholic University of Korea, Seoul 02451, Republic of Korea; ly1273701992@gmail.com³ School of Electronics and Information Engineering, Harbin Institute of Technology, Harbin 150080, China; lyzizizi@163.com

* Correspondence: hitfenglei@hit.edu.cn

Abstract

Automatic Music Transcription (AMT) plays a fundamental role in Music Information Retrieval (MIR) by converting raw audio signals into symbolic representations such as MIDI or musical scores. Despite advances in deep learning, accurately transcribing piano performances remains challenging due to dense polyphony, wide dynamic range, sustain pedal effects, and harmonic interactions between simultaneous notes. Existing approaches using convolutional and recurrent architectures, or autoregressive models, often fail to capture long-range temporal dependencies and global harmonic structures, while conventional Vision Transformers overlook the anisotropic characteristics of audio spectrograms, leading to harmonic neglect. In this work, we propose SpectTrans, a novel piano transcription framework that integrates a Spectral Gating Network with a multi-head self-attention Transformer to jointly model spectral and temporal dependencies. Latent CNN features are projected into the frequency domain via a Real Fast Fourier Transform, enabling adaptive filtering of overlapping harmonics and suppression of non-stationary noise, while deeper layers capture long-term melodic and chordal relationships. Experimental evaluation on polyphonic piano datasets demonstrates that this architecture produces acoustically coherent representations, improving the robustness and precision of transcription under complex performance conditions. These results suggest that combining frequency-domain refinement with global temporal modeling provides an effective strategy for high-fidelity AMT.

Keywords: automatic music transcription (AMT); pedal transcription; Spectral Transformers



Academic Editor: Ping-Feng Pai

Received: 31 December 2025

Revised: 24 January 2026

Accepted: 29 January 2026

Published: 3 February 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Automatic Music Transcription (AMT) refers to the process of converting raw audio signals into symbolic representations, such as MIDI events or musical scores, and constitutes a core research topic within Music Information Retrieval (MIR) [1]. Its objective extends beyond conventional Multi-Pitch Estimation (MPE) to the high-fidelity reconstruction of note-level attributes, including precise onset and offset times, pitch values, and velocities.

Early transcription practices relied heavily on oral tradition and manual notation, which were inherently limited by human subjectivity and insufficient temporal and pitch precision. As musical compositions have become increasingly complex and stylistically diverse, manual transcription has proven inadequate to meet the demands of contemporary musicological research and industrial applications. Recent advances in polyphonic

piano transcription have been driven by deep convolutional neural networks (CNNs) and recurrent neural networks (RNNs) [2–8]. In particular, joint onset-and-frame prediction frameworks have substantially improved note-level transcription accuracy. References [3,9] introduced an adversarial training strategy that formulates transcription as an image-to-image translation task and employs a fully convolutional discriminator, thereby alleviating the limitation of conventional element-wise loss functions that fail to model inter-label probabilistic dependencies. Kong et al. [10] further proposed a high-resolution AMT system that enhances transcription precision by explicitly regressing accurate onset and offset timings for piano notes. Despite these advances, autoregressive (AR) models suffer from inherent drawbacks due to their reliance on narrow sequential processing windows [11], which makes them vulnerable to error accumulation and exposure bias. From a musico-logical perspective, such models also struggle to capture the global structural hierarchy of musical compositions, including long-range rhythmic symmetries and recurring thematic patterns. Moreover, a fundamental challenge in polyphonic transcription lies in the harmonic coupling of simultaneously sounding notes. In Western tonal music, consonant intervals such as perfect fifths and major thirds generate strongly intertwined frequency components. Traditional convolutional or sequential architectures, which are primarily designed to model local spatial or temporal patterns, often fail to effectively disentangle these complex cross-frequency interactions.

Transformer-based architectures leveraging Multi-Head Self-Attention (MHSA) have recently emerged as a powerful alternative for AMT [12], owing to their superior capability in modeling long-range dependencies. Among these, Vision Transformers (ViTs) have been adapted to the AMT domain. However, their original design is tailored to natural images, which are characterized by translation-invariant spatial features. In contrast, audio spectrograms exhibit pronounced anisotropy, wherein the vertical axis encodes pitch-dependent harmonic structures, while the horizontal axis represents temporal evolution and rhythmic dynamics. Standard ViT patching strategies treat low-frequency fundamentals and high-frequency overtones uniformly, thereby disregarding the physical constraints imposed by the harmonic series. As a result, the direct application of ViTs to spectrogram representations often leads to harmonic negligence, a phenomenon that fundamentally contradicts the principles of musical acoustics. To address this limitation, recent studies have proposed harmonic-aware Transformer architectures that incorporate pre-designed harmonic mask matrices [13] or specialized harmonic attention mechanisms [14] to better capture spectral structures. Nevertheless, these approaches predominantly rely on rigid harmonic masks, i.e., hard-coded priors derived from idealized acoustic formulations. Such fixed associations lack the flexibility required to model the spectral variability, timbral diversity, and inharmonicity commonly present in real-world musical recordings.

In contrast, this paper proposes SpectTrans, which, to the best of our knowledge, represents the first integration of an Acoustic Spectrum Gating Network (A-SGN) into the piano transcription domain. Unlike previous methods based on fixed harmonic ratios, SpectTrans adopts a fully data-driven and learnable gating strategy. This design enables not only the modeling of harmonic relationships, but also the adaptive identification and suppression of complex acoustic background noise and non-harmonic interference. Specifically, in the shallow layers, the A-SGN embedded within the As-Attention module employs a learnable frequency-domain gating mechanism that emulates physically meaningful spectral filters. This structure explicitly encodes the logarithmic harmonic relationships inherent in piano timbres, thereby facilitating more effective separation of overlapping notes in polyphonic scenarios. In the deeper layers, an MHSA mechanism is employed to capture long-range temporal dependencies, enabling the modeling of complex melodic trajectories and chordal structures. Experimental results demonstrate that the proposed

framework significantly improves both the robustness and accuracy of piano transcription across diverse and challenging performance conditions.

2. Related Works

2.1. Automatic Piano Transcription

AMT is a cornerstone task in Music Information Retrieval (MIR), aiming to map complex raw audio signals into structured symbolic representations. Owing to the piano's wide dynamic range, dense polyphony, and additional acoustic complexity introduced by sustain pedal resonance, AMT remains a particularly challenging problem within the field. Early transcription paradigms primarily relied on frame-level classification schemes. However, such approaches are inherently constrained by temporal resolution determined by the hop size, leading to quantization artifacts and training instabilities caused by label misalignment. To overcome these limitations, Kong et al. [10] proposed a high-resolution regression framework that achieves sub-frame temporal precision by directly predicting absolute onset and offset time deviations. In a complementary direction, Wei et al. [9] introduced a dual-decoder architecture tailored for streaming transcription, employing causal attention mechanisms to decouple onset and offset estimation and thereby preserve logical note consistency in real-time scenarios. Convolutional Neural Networks (CNNs) have become the de facto standard for front-end feature extraction in AMT systems. Nevertheless, their inherently local receptive fields limit their ability to capture harmonic correlations spanning multiple octaves. To address long-range dependencies, sequential modeling architectures based on Recurrent Neural Networks (RNNs) [15] and, more recently, Transformers [16] have gained increasing attention. A notable trend within this line of research is the incorporation of music-informed inductive biases. For instance, Wang et al. [13] proposed a harmonic-aware frequency attention mechanism that employs explicit masking strategies to isolate energy distributions across overtone series. Similarly, Wu et al. [14] introduced Harmonic Attention to strengthen the modeling of acoustic priors inherent in musical signals. To mitigate the computational burden associated with long-sequence modeling, Wei and Yoshii [17] leveraged sparse attention mechanisms and hierarchical pooling strategies, thereby transcending sliding-window constraints and enabling more effective modeling of large-scale musical structures and sustained note envelopes. In our previous work [18], we further explored these relationships through a graph-theoretic perspective by proposing a Multi-Scale Graph Attention Network [19], in which musical notes are treated as graph nodes to capture the structural intricacies of piano textures. Despite these advances, a persistent bottleneck remains in the difficulty of adaptively distilling salient harmonic content from low-level acoustic representations while effectively suppressing complex background noise.

To bridge this gap, we propose SpectTrans, a novel architecture that fuses frequency-domain gating with temporal attention. Unlike traditional models confined to local spatial or temporal operations, SpectTrans introduces a Spectral Gating Network. By applying a Real Fast Fourier Transform, we project latent CNN features into the frequency domain for global refinement. This design leverages the global receptive field of the Fourier transform to perform adaptive filtering and noise suppression at an early stage. This “spectral shaping followed by temporal association” strategy provides the subsequent Transformer modules with acoustically consistent representations, significantly bolstering transcription robustness in extreme polyphonic conditions.

2.2. Spectral Transformers

The emergence of Spectral Transformers has been largely motivated by a broader shift toward alternative token-mixing paradigms, exemplified by architectures such as

MLP-Mixer. Rather than relying on the $O(N^2)$ computational complexity inherent to spatial self-attention, these approaches exploit the global receptive field and computational efficiency of frequency-domain operations to enhance feature representations. A seminal contribution in this direction is FNet [20], which demonstrated that replacing self-attention with parameter-free one-dimensional Fourier transforms can achieve competitive performance while substantially reducing inference latency. Building upon this idea, GFNet [21] introduced learnable frequency-domain filters, effectively reformulating token mixing as a depth-wise global convolution. By leveraging the global receptive field of the Fourier transform, these methods capture long-range spatial dependencies that are often inaccessible to local convolutional kernels. Beyond discrete spectral transformations, Guibas et al. [22] advanced this paradigm by framing token mixing through operator learning. Using Fourier Neural Operators (FNOs) [23], their models achieve resolution-invariant generalization by treating the Fourier transform as a functional operator. FNOs typically retain low-frequency modes and discard high-frequency components, performing rigid spectral truncation that limits flexibility for tasks requiring fine-grained frequency tuning.

Recent theoretical studies, such as [24], explore the mathematical foundations of attention, interpreting it as kernel regression. Wave-ViT [25] further enhances Transformer models by incorporating wavelet transforms for multiscale feature modeling and down-sampling in attention blocks. In piano transcription, harmonic-aware models impose fixed harmonic masks based on idealized acoustic formulas, enforcing attention to predefined harmonic positions. While effective, this rigid approach limits flexibility, as it relies on static frequency templates that fail to adapt to real-world spectral variations and non-harmonic interference in complex polyphonic music.

In contrast, the A-SGN presented here offers a flexible, data-driven frequency selection mechanism. Unlike harmonic-aware models, SGN does not rely on fixed templates but uses learnable gating weights and biases to adaptively modulate the frequency spectrum. This dynamic gating enables the model to selectively preserve relevant frequency components, suppress non-harmonic noise, and capture a broader range of spectral features, leading to improved robustness in polyphonic piano transcription.

3. Methods

3.1. Acoustics Spectral Gating Network

As illustrated in Figure 1b, the Acoustic Spectral Gating Network (A-SGN) can be viewed as a trainable global filter designed to capture the inherent periodicity and harmonic constraints of the musical signal within the input window. Considering periodic boundary conditions and element-wise spectral modulation, applying spectral gating in the frequency domain is mathematically equivalent to performing a circular convolution in the time domain, followed by bias addition. Accordingly, for a frequency-domain input $\mathcal{X}$, the acoustic spectral gating operation is equivalent to a circular convolution in the time domain, which can be formulated as

$$\mathcal{X} \cdot \mathcal{W} + \mathcal{B} = \mathcal{F}(X * W + B), \quad (1)$$

where $\mathcal{W} \in \mathbb{C}^{D \times D}$ and $\mathcal{B} \in \mathbb{C}^D$ denote the complex-valued weight matrix and bias vector, respectively. $W \in \mathbb{R}^D$ and $B \in \mathbb{R}^D$ denote the inverse-DFT counterparts of $\mathcal{W}$ and $\mathcal{B}$, respectively. The operator “ $*$ ” denotes circular convolution. For a multi-layer network architecture, the single-layer frequency-domain gating operation described in Equation (1) can be generalized to a hierarchical recursive form. Specifically, the frequency-domain linear gating is employed as the basic transformation unit at each layer, followed by a nonlinear activation function $\sigma(\cdot)$, thereby constructing a multi-layer spectral gating

network. Consequently, the output of the ℓ -th layer of the network can be recursively expressed as

$$h^\ell = \sigma(h^{\ell-1}\mathcal{W}^\ell + \mathcal{B}^\ell), h^0 = \mathcal{X}. \quad (2)$$

where h^ℓ denotes the output of the ℓ -th layer. To fully leverage the structural advantages of complex-valued representations, the linear transformation is explicitly decomposed into its constituent real and imaginary components. Specifically, the complex-valued weight matrix denotes as $W^\ell = W_R^\ell + jW_I^\ell$ and the bias vector denotes as $\mathcal{B}^\ell = \mathcal{B}_R^\ell + j\mathcal{B}_I^\ell$. By applying the algebraic rules of complex multiplication, the layer-wise computation is formulated as follows:

$$\text{Re}(h)^\ell = \text{Re}(h^{\ell-1})W_R^\ell - \text{Im}(h^{\ell-1})W_I^\ell + \mathcal{B}_R^\ell, \quad (3)$$

$$\text{Im}(h)^\ell = \text{Re}(h^{\ell-1})W_I^\ell + \text{Im}(h^{\ell-1})W_R^\ell + \mathcal{B}_I^\ell, \quad (4)$$

where h^ℓ represents the pre-activation state. To maintain the complex-valued nature of the features, a component-wise non-linear activation $\sigma(\cdot)$ is applied, yielding the final output:

$$h^\ell = \sigma(\text{Re}(h)^\ell) + j\sigma(\text{Im}(h)^\ell). \quad (5)$$

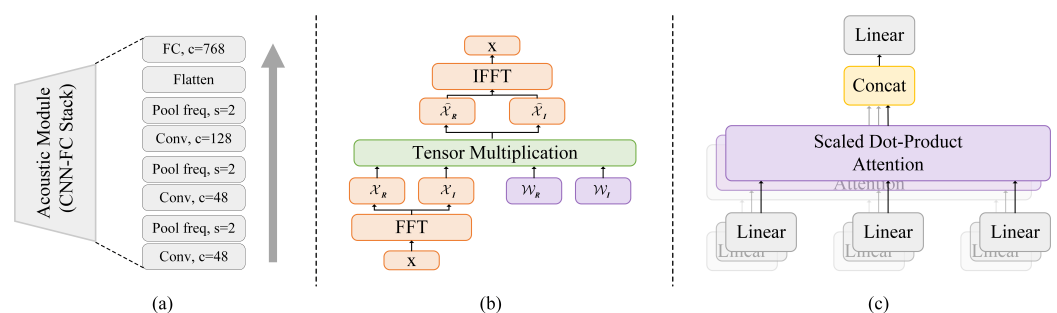


Figure 1. The architecture of the proposed model components. (a) The Acoustic Module consisting of a CNN-FC stack for feature extraction. (b) The Spectral Layer utilizing Fast Fourier Transform (FFT) and learnable weights for frequency-domain processing. (c) The Multi-head Attention mechanism for capturing contextual dependencies.

By reformulating the spectral gating mechanism through a complex-valued decomposition, the model explicitly characterizes the cross-correlation between the real and imaginary components. This approach effectively expands the hypothesis space of the gating operation, enabling the capture of more sophisticated frequency-domain dependencies. Such a formulation is instrumental in capturing the subtle inter-harmonic interference and sympathetic resonance inherent in piano tones. By precisely modeling these cross-component interactions, the network can distinguish between overlapping overtones that would otherwise lead to representation aliasing in a purely real-valued space. Operationally, the proposed A-SGN formulates the spectral gating mechanism as a complex-valued linear transformation, which is mathematically equivalent to a global circular convolution in the spatial domain. This formulation effectively mitigates the computational overhead associated with conventional spatial-domain convolutions. Specifically, whereas a global convolution in the spatial domain incurs a computational complexity of $O(N^2)$, the A-SGN transforms the operation into an element-wise modulation in the frequency domain, thereby reducing the complexity to a quasi-linear scale of $O(N \log N)$ via the Fast Fourier Transform (FFT). Subsequently, the Inverse Fast Fourier Transform (IFFT) is applied to map the processed signals from the frequency domain back to the latent space. Therefore, the

acoustic spectral gating operation can be effectively interpreted as a global convolution performed on the input features.

3.2. Acoustics Spectral Attention

In the traditional attention mechanism, the input sequence X undergoes parallel projections into specialized latent spaces. Specifically, the Query and Key components are mapped to different latent spaces using weight matrices W_Q and W_K , respectively, capturing the contextual dependencies between input tokens, primarily through the computation of scaled dot-product attention scores. These scores are then used to compute the attention weights A_{ij} , which represent the attention of token i to token j .

At the same time, the Output-Value (OV) circuit applies the transformation matrix W_V to generate value vectors v_i , which encode information. The final output R_i is a linear weighted sum of these value vectors, based on the attention distribution A , as expressed by:

$$R_i = \sum_j A_{ij} v_j, \quad (6)$$

where v_j denotes the value vector of token j , and A_{ij} represents the attention weight, indicating the importance of token j to token i .

However, while attention mechanisms excel at capturing long-range dependencies, they are less efficient in modeling periodic signals compared to spectral modules. To address this, we propose integrating the attention output R_i with the spectral result to enable complementary learning across different domains. The spectral result is derived through the A-SGN (Acoustic Spectral Graph Network) module, which interprets the element-wise modulation in the frequency domain as a kernel integral operator $\mathcal{K}$ in the spatial domain. This perspective allows A-SGN to be viewed as a learnable neural spectral kernel κ_ϕ , which learns continuous input-output mappings.

Specifically, the evolution of the latent state $v_\ell \mapsto v_{\ell+1}$ incorporates the non-local kernel $\mathcal{K}$ and a non-linear activation function $\sigma(\cdot)$, with the update rule given by:

$$v_t(x) \mapsto v_{t+1}(x) = \sigma((\mathcal{K}(a; \phi) v_t)(x)), \quad \forall x \in D, \quad (7)$$

where the kernel integral operator $\mathcal{K}$ is defined as:

$$(\mathcal{K}(a; \phi) v_t)(x) := \int_D \kappa(x, y, a(x), a(y); \phi) v_t(y) dy. \quad (8)$$

Here, $v_t(y)$ is the value vector at time step t , and D represents the input domain.

In this framework, κ_ϕ represents a neural network parameterized by $\phi \in \Theta_{\mathcal{K}}$, which allows the kernel integral operator to be implemented as a convolution operator in the Fourier domain. The A-SGN module captures the spectral characteristics of the input signal in the frequency domain and combines this with the temporal attention mechanism.

Finally, the Acoustic Spectral Attention mechanism integrates the spectral transformation and attention mechanism, expressed as:

$$v_{t+1}(x) := \mathcal{R}_t + \sigma((\mathcal{K}(a; \phi) v_t)(x)), \quad \forall x \in D, \quad (9)$$

where $\mathcal{R}_t$ represents the accumulation from the attention mechanism, v_t is the previous representation, and $\sigma(\cdot)$ is a non-linear activation function. The kernel function κ , parameterized by a and ϕ , acts on the previous representation v_t . Through this integration, the local focus of attention and the global reach of the spectral kernel are effectively combined, enabling the model to learn periodic and aperiodic signal transformations in complex acoustic signals.

3.3. SpectTrans

Based on the proposed Acoustics Spectral Gating Network (A-SGN) and the Acoustics Spectral Attention mechanism, we construct the overall transcription architecture, termed SpectTrans. SpectTrans is designed as a unified frequency–time domain hybrid modeling framework that jointly exploits spectral-domain global modeling and time-domain contextual attention. Specifically, as illustrated in Figure 2, SpectTrans first employs the Acoustic Module shown in Figure 1a to extract latent representations, providing a compact and informative embedding space for subsequent modeling. Each convolutional block consists of two convolutional layers with a kernel size of 3×3 . To ensure training stability and enhance the nonlinear representation capacity of the system, Batch Normalization [26] and ReLU activation [27] are applied immediately after each linear convolution operation. The extracted latent features are then processed by the SpectTrans backbone, which integrates the Acoustics Spectral Attention mechanism with the standard Transformer attention architecture, as shown in Figure 1c. The spectral pathway implemented by the A-SGN performs global frequency-domain modulation, enabling the model to capture intrinsic harmonic structures and long-range spectral correlations inherent in piano acoustics. In parallel, the standard attention mechanism models long-range temporal dependencies, contextual relationships across frames, and event-level dynamics in the time domain. This dual-pathway design enables complementary learning in the frequency and temporal domains, allowing SpectTrans to jointly model global periodic structures and long-range temporal dependencies. To further enhance onset detection performance, considering that note velocity provides informative cues for onset localization, the outputs of the velocity regression submodule and the onset regression submodule are concatenated along the frequency dimension. This concatenated representation is then fed into a cross-attention layer to compute the final onset predictions. Similarly, the outputs of the onset regression and offset regression submodules are concatenated and used as the input to a cross-attention layer to generate the final frame-wise predictions. The final model output is denoted as $Y_q \in \mathcal{R}_{L \times K}$, where $q \in \{frame, onset, offset, velocity\}$ represents the piano roll predictions over L frames and K piano pitches for each transcription subtask, including frame-level pitch estimation, onset detection, offset detection, and velocity estimation. Binary cross-entropy loss is employed for all subtasks.

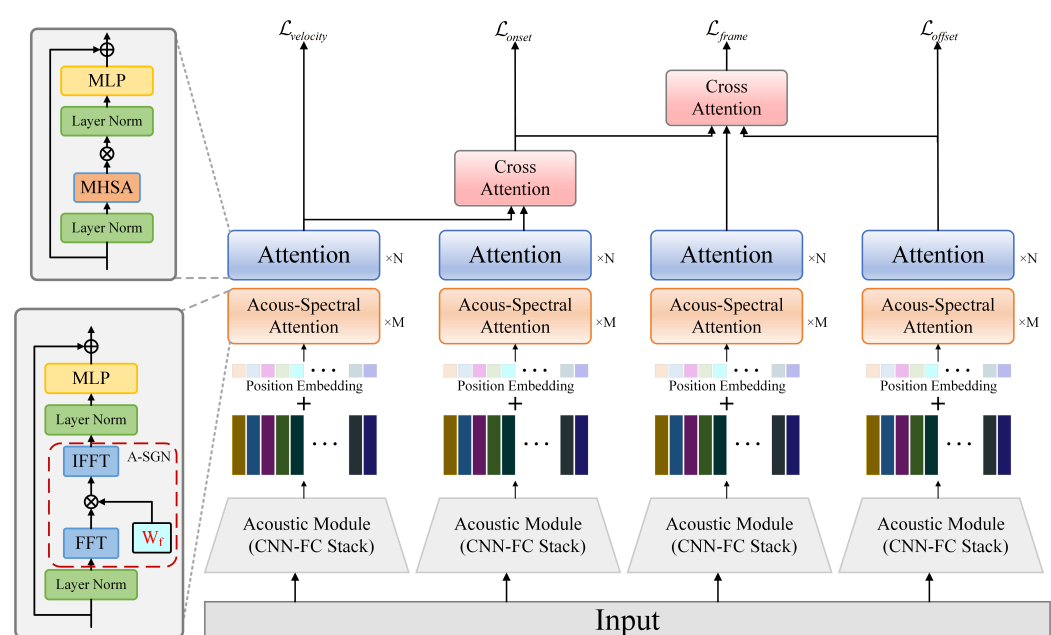


Figure 2. The overall framework of our proposed SpectTrans.

The spectral gating component introduces a structured global convolutional prior through frequency-domain modulation, enhancing the model's robustness to harmonic interference and spectral aliasing. Meanwhile, the Transformer component provides flexible sequence modeling capability, enabling the network to capture note-level temporal evolution, onset–offset relationships, and cross-frame dependencies, which are essential for accurate piano transcription. The joint optimization of these two components allows SpectTrans to achieve a balanced representation that is both spectrally structured and temporally coherent. Therefore, SpectTrans can be interpreted as a hybrid frequency–time domain architecture, in which frequency-domain global modeling and attention-based temporal contextual learning are tightly coupled. This design effectively overcomes the limitations of purely spectral models or purely attention-based models, enabling effective modeling of complex acoustic signals with strong harmonic structures and long-range temporal dependencies.

4. Results

4.1. Datasets and Evaluation Metrics

Experimental evaluation was conducted on the widely adopted MAESTRO (v1.0.0) piano dataset [2], which contains approximately 200 h of high-fidelity recordings of live piano performances paired with precisely aligned MIDI annotations. These MIDI files provide detailed information on note events, including key-strike velocities and sustain pedal positions. All experiments strictly follow the official dataset splits for training, validation, and testing to ensure fair and reproducible comparisons. During audio preprocessing, the raw waveforms were resampled to 16 kHz using the librosa library [28] and segmented into fixed-length 8-second audio clips. MIDI data parsing was performed using the pretty_midi library [29]. Each note event was independently projected onto four separate piano-roll representations corresponding to frame activation, onset, offset, and velocity, respectively, providing a comprehensive symbolic encoding for supervised training. For a quantitative assessment of the effectiveness of the proposed method, we introduce three evaluation metrics, encompassing Precision, Recall and F1 score [30], to evaluate the system performance at both the frame-level and note-level metrics. The formulas are as follows:

$$Precision = \frac{TP}{TP + FP} \quad (10)$$

$$Recall = \frac{TP}{TP + FN} \quad (11)$$

$$F1 = \frac{2Precision \times Recall}{Precision + Recall} \quad (12)$$

where TP is the number of true positives, FP is the number of false positives and FN is the number of false negatives.

4.2. Experiment Setup

Following the established protocol, we employ high-resolution regression for the precise estimation of onset and offset times within their respective branches. The training process is optimized using the Adam [31] optimizer with a mini-batch size of 8. We maintain a learning rate of 0.0005 for 300k steps, incorporating an early stopping strategy to ensure convergence. Model training was conducted on two NVIDIA GeForce RTX 5090 GPUs (Nvidia, Santa Clara, CA, USA).

4.3. Evaluation Results of Piano Transcription

Table 1 presents the performance of our proposed piano transcription method on the MAESTRO V1.0.0 test set. We compare our method with several representative methods, including Onsets and Frames [9], Adversarial Onsets and Frames [3], S2S [32], PAR [33], and HRT [10]. The method in [3] extends [9] by incorporating adversarial training and is likewise optimized with a quantization-based objective. S2S employs a generic encoder–decoder Transformer architecture with standard decoding strategies to directly translate spectrogram inputs into MIDI-like event sequences. PAR constructs a real-time piano transcription system by introducing a FiLM-based frequency modulation layer, a pitch-wise LSTM, and enhanced autoregressive contextual modeling. HRT proposes a high-resolution piano transcription framework that regresses precise note onset and offset times, thereby surpassing traditional frame-level detection approaches. To ensure a fair comparison between our proposed approach and prior methods, we adopt consistent evaluation criteria. Specifically, a tolerance of 50 ms is used for onset evaluation, while offset evaluation applies the same 50-millisecond tolerance together with an offset ratio of 0.2. For velocity evaluation, a tolerance threshold of 0.1 is employed. After scaling and normalizing velocities to the range [0, 1], an estimated note is regarded as correct if its velocity falls within the specified tolerance of the corresponding reference note. Compared with baseline, our proposed SpectTrans improves the frame F1 score from 82.7% to 89.2% and note F1 score from 96.1% to 96.6% and improves the note F1 score evaluated with offset from 89.6% to 90.3%. These results demonstrate that our proposed method, by combining a spectral gating network with a multi-head self-attention mechanism, effectively models both spectral characteristics and temporal dependencies, thereby achieving competitive performance.

Table 1. Transcription results evaluated on the test set of MAESTRO V1 dataset. Results in bold indicate the method proposed in this paper. “-” indicates that the corresponding result is not available.

Methods	Frame			Note			Note w/ Offset		
	P (%)	R (%)	F1 (%)	P (%)	R (%)	F1 (%)	P (%)	R (%)	F1 (%)
Onsets & frames	-	-	-	92.6	98.3	95.3	78.2	83.0	80.5
Adversarial	-	-	-	93.2	98.1	95.6	79.3	83.5	81.3
S2S	-	-	-	-	-	96.0	-	-	83.5
PAR _{compact}	-	-	-	94.9	97.2	96.0	83.7	85.7	84.7
HRT (baseline)	87.4	90.1	82.7	98.3	97.4	96.1	92.7	88.9	89.6
SpectTrans	88.1	90.5	89.2	98.0	95.3	96.6	93.0	88.0	90.3

The cross-domain evaluation results on the MAPS dataset are presented in Table 2. We strictly follow the standard cross-domain evaluation protocol [10], where the model is trained on the MAESTRO dataset without any data augmentation and directly tested on the MAPS dataset. For comparison, we also evaluate HRT under the same conditions. As shown in Table 2, the proposed SpectTrans achieves competitive performance, attaining an F1 score of 77.6% in frame-level detection, an F1-score of 81.6% for note onset detection and 56.7% for note onset with offset. These results provide strong evidence that SpectTrans demonstrates robust generalization capability and resilience to variations in piano timbre and non-ideal recording environments.

Table 2. Cross-domain evaluation results on the MAPS Dataset. Results in bold indicate the method proposed in this paper.

Methods	Frame			Note			Note w/ Offset		
	P (%)	R (%)	F1 (%)	P (%)	R (%)	F1 (%)	P (%)	R (%)	F1 (%)
HRT (baseline)	85.2	80.2	82.5	78.9	87.4	82.8	54.0	59.9	56.2
SpectTrans	85.3	79.7	82.7	77.6	87.3	81.6	54.7	60.3	56.7

4.4. Evaluation Results of Pedal Piano Transcription

The actual duration of a musical note depends not only on the release of the key (note offset) but also on the state of the sustain pedal. When the pedal is pressed, the note continues to sound even after the finger leaves the keyboard. Therefore, following existing evaluation protocols, we further report the experimental results of pedal transcription, as shown in Table 3. Specifically, with a pedal onset tolerance of 50 ms, the proposed SpectTrans achieved an F1 score of 94.80% in event-based evaluation; with a pedal onset tolerance of 50 ms and an offset ratio of 0.2, SpectTrans also achieved an F1 score of 79.76% in event-based evaluation. These results demonstrate that by simultaneously performing high-resolution regression modeling of both notes and pedals, our method can more accurately characterize the acoustic boundaries of notes, thus achieving high-resolution piano transcription.

Table 3. Pedal Transcription evaluated on the test set of MAESTRO V1 dataset. Results in bold indicate the method proposed in this paper.

Methods	Frame			Event			Event w/ Offset		
	P (%)	R (%)	F1 (%)	P (%)	R (%)	F1 (%)	P (%)	R (%)	F1 (%)
HRT	94.76	94.72	94.69	90.70	92.56	91.46	80.13	78.61	79.07
SpectTrans	95.04	94.91	94.91	97.42	92.37	94.80	81.84	77.66	79.67

4.5. Ablation Studies

In our architecture, the A-SGN is designed as a performance-enhancement module integrated into the Transformer backbone. To isolate and quantify its contribution, we conduct a controlled ablation study by comparing SpectTrans with standard attention and alternative spectral layer variants, including the Fourier Network (FN) and the discrete cosine transform (DCT), as summarized in Table 4. The results show that replacing standard attention with FN reduces both the parameter count (106.47 M vs. 125.37 M) and FLOPs (214.83 vs. 233.72), while achieving a frame-level F1-score of 89.3%. The DCT variant achieves a slightly higher frame-level F1-score of 89.1%, but at the cost of increased computational complexity (231.82 FLOPs). In contrast, SpectTrans, which incorporates the A-SGN, achieves the best frame-level F1-score of 89.2% with minimal computational cost (106.46 M parameters and 214.82 FLOPs). These results demonstrate that A-SGN provides a more effective spectral modeling mechanism than alternative spectral layers, yielding superior frequency-domain feature discrimination and noise suppression while preserving computational efficiency.

Table 4. Spectral layer variants. This table shows the ablation analysis of various spectral layers in SpectFormer architecture such as the FN, the FNO, and the A-SGN. We conduct this ablation study on the small-size networks in stage architecture. This indicates that FGN performs better than other kinds of networks. Results in bold indicate the method proposed in this paper.

Methods	Parameters	FLOPs	Frame		
			P (%)	R (%)	F1 (%)
Standard attention	125.37	233.72	-	-	-
FN	106.47	214.83	87.9	87.8	89.3
DCT	109.51	231.82	88.3	89.6	89.1
SpectTrans	106.46	214.82	88.1	90.5	89.2

4.6. Computational Efficiency

This subsection focuses on analyzing the computational efficiency of the proposed architecture. As reported in Table 4, although SpectTrans integrates a CNN-based acoustic front-end, a Spectral Gating Network, and a Transformer backbone, the overall model complexity remains low, with 106.46 M parameters and 214.82 FLOPS. Compared with the baseline model using standard self-attention (125.37 M parameters and 233.72 FLOPS), SpectTrans achieves both higher transcription accuracy and lower computational cost, indicating that performance gains are not achieved at the expense of efficiency.

This efficiency advantage originates from the structural properties of the spectral gating mechanism. Spectral gating operates through element-wise modulation in the frequency domain, which has linear computational complexity with respect to sequence length. In contrast, standard self-attention incurs quadratic complexity ($O(N^2)$) due to pairwise token interactions. As a result, SGN serves as a low-cost alternative to conventional attention, enabling efficient global information filtering without introducing significant computational overhead. This design demonstrates that structurally richer architectures can achieve improved modeling capacity without proportional increases in computational complexity.

4.7. Visualization of Frame Activations

Figure 3 illustrates the output of the piano transcription system when processing complex piano recordings. The first row shows the log-Melogram of the input audio segment, serving as the basis for the model to extract time-frequency features. The second row shows the frame-level note activation probabilities, reflecting the sustained state of the notes over time. The third row shows the regression output of the note initiation point, used to capture the precise moment of note attack. The fourth row shows the regression output of the note termination point, used to identify the precise moment of note release. The fifth row shows the activation curve of the sustain pedal, recording the continuous state of the pedal being pressed and released during performance. This visualization outputs provides qualitative evidence of the SpectTrans framework's superior performance in high-fidelity piano transcription. As shown in the onset and offset panels, the proposed methods produces sharp and localized activation peaks which demonstrate that the Spectral Gating Network effectively mitigates harmonic interference and non-stationary noise through frequency-domain refinement. This specialized gating process addresses the anisotropic nature of audio spectrograms and prevents the “harmonic neglect” commonly found in standard Vision Transformers by adaptively filtering overlapping partials.

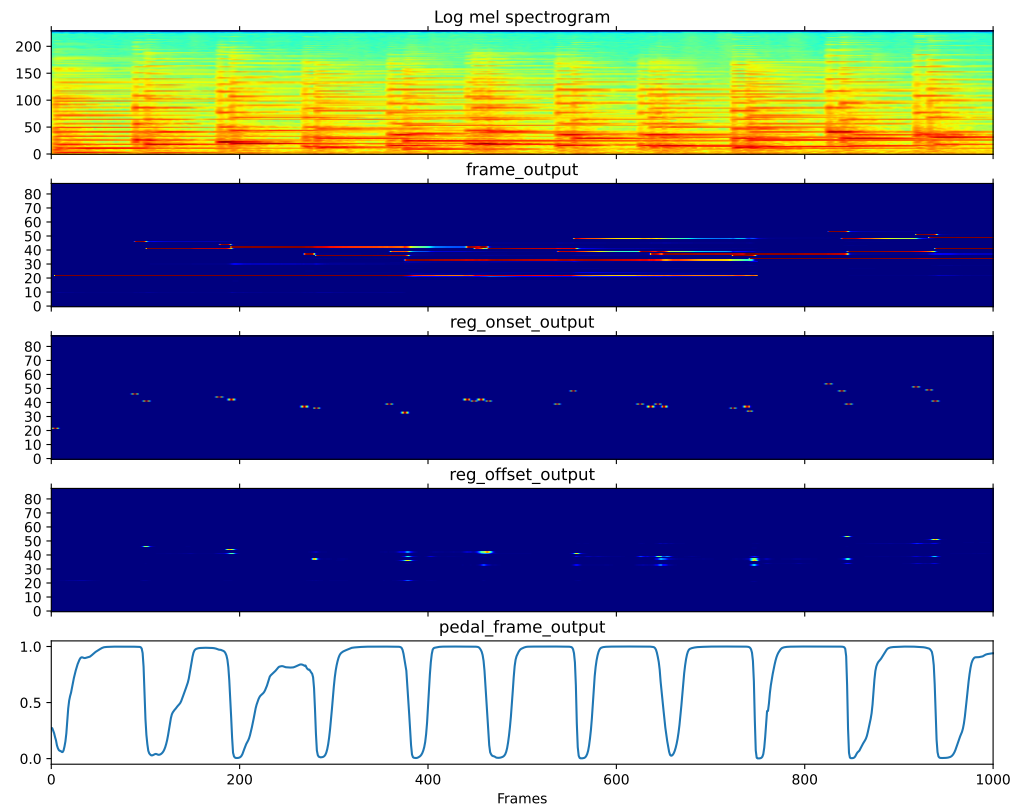


Figure 3. Visualization of the transcription outputs for a piano performance segment. From top to bottom, the panels show the log-Mel spectrogram of the input audio, the frame-wise probability of note activation, the onset regression indicating note attack moments, the offset regression identifying note release moments, and the pedal activation curve representing the sustain pedal state. For the spectrogram and regression-related outputs, warmer colors correspond to higher energy or higher prediction confidence. The horizontal axis denotes time frames, and the overall temporal alignment demonstrates the model’s capability to accurately map symbolic music events to the underlying acoustic signal.

5. Error Analysis and Limitation

While SpectTrans demonstrates strong overall performance, a detailed error analysis highlights specific limitations under challenging acoustic conditions. One major issue arises in highly polyphonic passages, where spectral masking can obscure overlapping harmonics. Although the A-SGN effectively refines harmonic content, it occasionally struggles to disentangle closely overlapping partials. Additionally, in passages with extensive sustain pedal use, temporal boundaries become less distinct, indicating that global spectral filtering may unintentionally smooth fine-grained transient offsets. Furthermore, the current analysis primarily focuses on spectral and temporal features, leaving the model’s ability to leverage higher-level musical structures—such as phrasing, motifs, and harmonic progression—largely unexplored. Addressing these aspects could provide a deeper understanding of systematic errors and guide improvements in transcription accuracy.

6. Conclusions

In this paper, we introduce SpectTrans, a polyphonic piano transcription framework that integrates spectral analysis with long-range temporal modeling. Central to our approach is the Acoustic Spectral Gating Network (SGN), which employs learnable frequency-domain gating to emulate physical frequency filtering and encode the logarithmic harmonic structure inherent to piano timbres. Complementing this, an attention mechanism captures long-range temporal dependencies to improve note recognition over extended contexts.

Experimental results demonstrate that SpectTrans achieves effective transcription performance across diverse piano pieces. Beyond piano transcription, the proposed architecture is adaptable to other audio analysis tasks, including multi-instrument transcription, speech separation, and sound event detection. Future work may focus on task-specific adaptations and parameterized extensions to further enhance the model's flexibility and accuracy.

For future work, two directions are particularly promising. First, the model's capacity to exploit higher-level musical structure remains underexplored and could enhance transcription robustness, especially in complex polyphonic passages. Second, a more comprehensive error analysis framework is needed to systematically identify limitations and guide refinements. These avenues offer opportunities to extend the model's applicability and improve its precision in real-world music transcription tasks.

Author Contributions: R.C.: Formal analysis, supervision, funding acquisition, software, and writing—original draft. Y.L. (Yan Liang): Software, methodology, validation, and writing—original draft. L.F.: Investigation, formal analysis, and visualization. Y.L. (Yuanzi Li): Conceptualization, formal analysis, and writing—reviewing. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding

Data Availability Statement: The datasets are available within the link <https://adasp.telecom-paris.fr/resources/>, accessed on 1 October 2024.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AMT	Automatic Music Transcription
MIR	Music Information Retrieval
SGN	Spectral Gating Network
A-SGN	Acoustic Spectral Gating Network
MHSA	Multi-Head Self-Attention
FFT/IFFT	Fast Fourier Transform/Inverse FFT
CNN	Convolutional Neural Network
RNN	Recurrent Neural Network
MLM	Music Language Model
FNO	Fourier Neural Operator

References

1. Mesaros, A.; Heittola, T.; Virtanen, T.; Plumbley, M.D. Sound event detection: A tutorial. *IEEE Signal Process. Mag.* **2021**, *38*, 67–83. [CrossRef]
2. Hawthorne, C.; Stasyuk, A.; Roberts, A.; Simon, I.; Huang, C.Z.A.; Dieleman, S.; Elsen, E.; Engel, J.; Eck, D. Enabling factorized piano music modeling and generation with the MAESTRO dataset. In Proceedings of the International Conference on Learning Representations (ICLR), New Orleans, LA, USA, 6–9 May 2019.
3. Kim, J.; Bello, J.P. Adversarial learning for improved onsets and frames music transcription. In Proceedings of the 20th International Society for Music Information Retrieval Conference (ISMIR), Delft, The Netherlands, 4–8 November 2019; pp. 670–677.
4. Kelz, R.; Böck, S.; Widmer, G. Deep polyphonic ADSR piano note transcription. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 246–250.
5. Kelz, R.; Dorfer, M.; Korzeniowski, F.; Böck, S.; Arzt, A.; Widmer, G. On the potential of simple framewise approaches to piano transcription. In Proceedings of the 17th International Society for Music Information Retrieval Conference (ISMIR), New York, NY, USA, 7–11 August 2016; pp. 475–481.
6. Kwon, T.; Jeong, D.; Nam, J. Polyphonic piano transcription using autoregressive multi-state note model. In Proceedings of the 21st International Society for Music Information Retrieval Conference (ISMIR), Virtual, 11–16 October 2020; pp. 454–461.

7. Wang, Q.; Zhou, R.; Yan, Y. Polyphonic piano transcription with a note-based music language model. *Appl. Sci.* **2018**, *8*, 470. [CrossRef]
8. Böck, S.; Schedl, M. Polyphonic Piano Note Transcription with Recurrent Neural Networks. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Kyoto, Japan, 25–30 March 2012; pp. 121–125.
9. Hawthorne, C.; Elsen, E.; Song, J.; Roberts, A.; Simon, I.; Raffel, C.; Engel, J.; Oore, S.; Eck, D. Onsets and Frames: Dual-Objective Piano Transcription. In Proceedings of the 19th International Society for Music Information Retrieval Conference (ISMIR), Paris, France, 23–27 September 2018; pp. 50–57.
10. Kong, Q.; Li, B.; Song, X.; Wan, Y.; Wang, Y. High-Resolution Piano Transcription with Pedals by Regressing Onsets and Offsets. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2021**, *29*, 3707–3717. [CrossRef]
11. Xiong, J.; Liu, G.; Huang, L.; Wu, C.; Wu, T.; Mu, Y.; Yao, Y.; Shen, H.; Wan, Z.; Huang, J.; et al. Autoregressive models in vision: A survey. *arXiv* **2024**, arXiv:2411.05902. [CrossRef]
12. Wei, F.; Yoshii, K. Streaming Piano Transcription with Causal Attention and Dual Decoders. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Singapore, 22–27 May 2022; pp. 806–810.
13. Wang, Y.; Wu, J.; Zhang, L. Harmonic-Aware Frequency Attention for Piano Transcription. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Toronto, ON, Canada, 6–11 June 2021; pp. 301–305.
14. Wu, J.; Wang, Y.; Duan, Z. Harmonic Attention Networks for Music Transcription. In Proceedings of the 21st International Society for Music Information Retrieval Conference (ISMIR), Virtual, 11–16 October 2020; pp. 111–118.
15. Medsker, L.R.; Jain, L. Recurrent neural networks: Design and applications. *Neural Netw.* **2001**, *5*, 2.
16. Han, K.; Wang, Y.; Chen, H.; Chen, X.; Guo, J.; Liu, Z. A survey on vision transformer. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 87–110. [CrossRef] [PubMed]
17. Wei, F.; Yoshii, K. Hierarchical Sparse Attention for Long-Sequence Music Transcription. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Toronto, ON, Canada, 6–11 June 2021; pp. 716–720.
18. Cao, R.; Liang, Z.; Yan, Z.; Liu, B. DAFE-MSGAT: Dual-Attention Feature Extraction and Multi-Scale Graph Attention Network for polyphonic piano transcription. *Electronics* **2024**, *13*, 3939. [CrossRef]
19. Zhang, X.; Liu, Y.; Duan, Z. Multi-Scale Graph Attention Networks for Piano Transcription. In Proceedings of the 23rd International Society for Music Information Retrieval Conference (ISMIR), Bengaluru, India, 4–8 December 2022; pp. 410–417.
20. Lee-Thorp, J.; Ainslie, J.; Eckstein, I.; Ontañón, S. FNet: Mixing Tokens with Fourier Transforms. In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Seattle, WA, USA, 10–15 July 2022; pp. 213–223.
21. Rao, Y.; Zhao, W.; Zhu, Z.; Lu, J.; Zhou, J. Global Filter Networks for Image Classification. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 980–993.
22. Guibas, J.; Mardani, M.; Li, Z.; Tao, A.; Anandkumar, A.; Catanzaro, B. Efficient Token Mixing for Transformers via Adaptive Fourier Neural Operators. In Proceedings of the International Conference on Learning Representations (ICLR), Virtual, 3–7 May 2021. Available online: <https://openreview.net/forum?id=EXHG-A3jIM> (accessed on 24 January 2026).
23. Li, Z.; Kovachki, N.; Azizzadenesheli, K.; Liu, B.; Bhattacharya, K.; Stuart, A.; Anandkumar, A. Fourier Neural Operator for Parametric Partial Differential Equations. In Proceedings of the International Conference on Learning Representations (ICLR), Virtual, 3–7 May 2021; pp. 1–16.
24. Choromanski, K.; Likhoshesterov, V.; Dohan, D.; Song, X.; Gane, A.; Sarlos, T.; Hawkins, P.; Davis, J.; Mohiuddin, A.; Kaiser, L.; et al. Rethinking Attention with Performers. In Proceedings of the International Conference on Learning Representations (ICLR), Virtual, 3–7 May 2021; pp. 1–17.
25. Yao, T.; Pan, Y.; Li, Y.; Ngo, C.W.; Mei, T. Wave-ViT: Unifying Wavelet and Transformer for Visual Representation Learning. In *Computer Vision—ECCV 2022, Proceedings of the 17th European Conference, Tel Aviv, Israel, 23–27 October 2022*; Springer: Cham, Switzerland, **2022**; pp. 328–345.
26. Ioffe, S.; Szegedy, C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In Proceedings of the International Conference on Machine Learning (ICML), Lille, France, 6–11 July 2015; pp. 448–456.
27. Nair, V.; Hinton, G.E. Rectified Linear Units Improve Restricted Boltzmann Machines. In Proceedings of the 27th International Conference on Machine Learning (ICML), Haifa, Israel, 21–24 June 2010; pp. 807–814.
28. Kong, Q.; Cao, Y.; Iqbal, T.; Wang, Y.; Wang, W.; Plumbley, M.D. PANNs: Large-scale pretrained audio neural networks for audio pattern recognition. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2020**, *28*, 2880–2894. [CrossRef]
29. Raffel, C.; Ellis, D.P.W. Intuitive analysis, creation and manipulation of MIDI data with *pretty_midi*. In Proceedings of the 15th International Conference on Music Information Retrieval Late Breaking and Demo Papers, Taipei, Taiwan, 27–31 October 2014.
30. Yacoub, R.; Axman, D. Probabilistic extension of precision, recall, and F1 score for more thorough evaluation of classification models. In Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems, Online, 20 November 2020; pp. 79–91.
31. Kingma, D.P. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.

32. Hawthorne, C.; Simon, I.; Swavely, R.; Manilow, E.; Engel, J. Sequence-to-sequence piano transcription with transformers. *arXiv* **2021**, arXiv:2107.09142.
33. Kwon, T.; Jeong, D.; Nam, J. Towards Efficient and Real-Time Piano Transcription Using Neural Autoregressive Models. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2024**, *32*, 5106–5116. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.