mathematics

MDPI

Article

PF2N: Periodicity–Frequency Fusion Network for Multi-Instrument Music Transcription

Taehyeon Kim [1], Man-Je Kim [2,*] and Chang Wook Ahn [1,*]

[1] AI Graduate School, Gwangju Institute of Science and Technology, Gwangju 61005, Republic of Korea; esteban12@gm.gist.ac.kr
[2] Convergence of AI, Chonnam National University, Gwangju 61186, Republic of Korea
* Correspondence: jaykim0104@jnu.ac.kr (M.-J.K.); cwan@gist.ac.kr (C.W.A.)

Abstract: Automatic music transcription in multi-instrument settings remains a highly challenging task due to overlapping harmonics and diverse timbres. To address this, we propose the Periodicity–Frequency Fusion Network (PF2N), a lightweight and modular component that enhances transcription performance by integrating both spectral and periodicity-domain representations. Inspired by traditional combined frequency and periodicity (CFP) methods, the PF2N reformulates CFP as a neural module that jointly learns harmonically correlated features across the frequency and cepstral domains. Unlike handcrafted alignments in classical approaches, the PF2N performs data-driven fusion using a learnable joint feature extractor. Extensive experiments on three benchmark datasets (Slakh2100, MusicNet, and MAESTRO) demonstrate that the PF2N consistently improves transcription accuracy when incorporated into state-of-the-art models. The results confirm the effectiveness and adaptability of the PF2N, highlighting its potential as a general-purpose enhancement for multi-instrument AMT systems.

Keywords: multi-instrument music transcription; joint feature extraction; combined frequency and periodicity

MSC: 68T07

Academic Editor: Ioannis Tsoulos

Received: 25 April 2025
Revised: 17 May 2025
Accepted: 21 May 2025
Published: 23 May 2025

Citation: Kim, T.; Kim, M.-J.; Ahn, C.W. PF2N: Periodicity–Frequency Fusion Network for Multi-Instrument Music Transcription. *Mathematics* **2025**, *13*, 1708. https://doi.org/10.3390/math13111708

1. Introduction

Automatic music transcription (AMT) [1] converts audio signals into symbolic representations, facilitating tasks in music information retrieval (MIR) such as music generation [2,3], classification [4,5], and source separation [6,7]. In particular, multi-instrument transcription remains challenging due to overlapping harmonics and diverse timbres, making pitch disentanglement difficult.

Despite significant advancements in single-instrument transcription using support vector machines [8,9], non-negative matrix factorizations [10–12], and deep learning [13–17], the task of assigning notes to their corresponding instruments remains a challenge. For instance, ReconVAT [18] focuses on transcription in low-resource environments but produces a single-track piano roll without instrument-level separation. To address this limitation, other works proposed unified models that jointly perform music source separation and transcription. Cerberus [19] introduced a multi-task model with a source separation head, a deep clustering head, and a transcription head. Additionally, ref. [20] explored a clustering-based approach to separate transcribed instruments. A zero-shot multi-task model was presented in ref. [21], integrating source separation, transcription, and synthesis within a unified framework. Further, inspired by instance segmentation and multi-object

detection in computer vision, ref. [22] proposed a self-attention-based U-Net to jointly model pitch and instrument attributes. Building on this trend, Jointist [23] extends the paradigm by unifying transcription, instrument recognition, and source separation through a conditionally trained framework, enabling flexible multi-instrument modeling.

A notable breakthrough in this area was achieved with MT3 [24], a T5-based model that maps input sequences to learned embeddings with fixed positional encodings. By standardizing test splits and evaluation metrics across datasets, MT3 established a strong baseline in the field. Building on this, YMT3 [25] introduces a training toolkit to facilitate the reproduction and training of T5-based multi-instrument transcription models. Perceiver TF [26] applies spectral cross-attention (SCA), channel-wise self-attention, and temporal self-attention, enabling the efficient processing of high-dimensional representations. Expanding on these strategies, YPTF [27] enhances Perceiver TF by incorporating a mixture of experts (MoE) within its encoder and a multi-channel decoder designed based on multi-T5 to process diverse instrument event tokens more effectively.

Although previous architectures showed notable advances, instrument-wise transcription remains challenging due to overlapping harmonics and reliance on frequency-domain features alone. To address these limitations, we propose the Periodicity–Frequency Fusion Network (PF2N), a compact and modular neural component designed to improve multi-instrument music transcription.

- **Neural Frequency–Periodicity Feature Fusion:** Inspired by traditional combined frequency and periodicity (CFP) methods [28], the PF2N reformulates spectral–cepstral alignment as a learnable pattern extraction mechanism.
- **Compact and Modular Architecture:** The proposed model is lightweight in both size and computational cost. It is also a modular neural component that can be integrated into various baseline architectures.
- **Fair Comparisons with Benchmarks:** Previous studies used different combinations of training datasets but evaluated them separately, leading to inconsistent benchmarks. To ensure fairness, we train and evaluate each model independently on individual datasets.

The remainder of this manuscript is organized as follows: Section 2 reviews related work, Section 3 details the PF2N, Section 4 covers the experimental setup and results, and Section 5 summarizes the key findings and future directions.

2. Related Works

2.1. Multi-Instrument Music Transcription

In the early stages of multi-instrument music transcription [19–22], the lack of sufficiently large datasets and standardized benchmarking protocols posed challenges for consistent model evaluation. Consequently, many studies either compared model performance against single-instrument baselines or conducted experiments on disparate datasets. The introduction of large-scale multi-instrument music datasets such as Slakh2100 [29], along with the MT3 [24] model's systematic evaluation across multiple datasets, helped to establish a unified benchmark for multi-instrument AMT.

Nevertheless, recent models, including Perceiver TF [26] and YPTF [27], differ significantly in dataset usage. For example, while MT3 obtains benchmark results by training and evaluating on each dataset independently, Perceiver TF and YPTF employ distinct data augmentation strategies and train on different mixtures of datasets. These discrepancies have resulted in non-uniform evaluation settings, making direct performance comparisons difficult. To address this, we adopt the standardized dataset splits introduced by

MT3 and evaluate all models on identical partitions of Slakh2100 [29], MusicNet [30], and MAESTRO [31], ensuring fair and consistent benchmarking.

2.2. Combined Frequency and Periodicity

The CFP [28] method demonstrated that both frequency-domain and quefrency-domain features are crucial for pitch estimation. Given an N-point segment of music audio signal $X \in \mathbb{R}^N$, these features are represented by the log-magnitude spectrogram $U(f) = [|\mathcal{F}_{N,h}x|] \in \mathbb{R}^{T\times F}$ and the cepstrogram $V(\frac{1}{f}) = \mathcal{F}_N^{-1}([|\mathcal{F}_{N,h}x|]) \in \mathbb{R}^{T\times F}$, where $\mathcal{F}$ is the N-point Discrete Fourier Transform (DFT) with a window function h. In this representation, a fundamental frequency (f_0) produces a harmonic series with peaks at integer multiples of f_0 in the $U(f)$ and a corresponding subharmonic series at integer divisions of f_0 in the $V(\frac{1}{f})$. CFP leverages this relationship to estimate pitch by aligning spectral and cepstral peaks, computing f_0 as $\hat{f}_0 = argmax_f(U(f)V(\frac{1}{f}))$. This alignment method highlights the importance of cepstral features in transcription models.

However, directly multiplying these representations is often insufficient, as the true f_0 does not always correspond to the first few local maxima of $U(f)V(1/f)$. To refine pitch selection, ref. [28] introduced a post-processing pipeline involving missing fundamental frequency detection, stacked harmonic selection, and post-processing. This inspired subsequent works such as [32–34], which extended CFP by incorporating harmonicity constraints [32], applying patch-based CNNs to spectro-cepstral inputs for vocal melody extraction [33], and leveraging generalized cepstral features as input channels to deep learning models to enhance pitch saliency and suppress harmonic interference [34].

While previous methods enhanced CFP representations through post-processing or task-specific feature engineering, a key limitation remains. In the case of stacked harmonics, pitch selection is based solely on the geometric mean of peak magnitudes, which lacks a solid theoretical foundation. To overcome this, Multi-Layered CFP (ML-CFP) [35] extends CFP by introducing multi-layered frequency-domain and quefrency-domain features, leveraging partial cepstra extracted via liftering to enhance robustness against harmonic interference. Despite such refinements, these methods still rely on handcrafted spectral–cepstral alignment, which makes them less flexible. Our work departs from this paradigm by embedding CFP directly into a neural architecture, enabling adaptive and data-driven alignment between spectral and cepstral domains.

3. Proposed Method: PF2N

To address the limitations of traditional CFP methods, we propose the PF2N, a learnable model that adaptively extracts, enhances, and fuses both representations. A shared convolutional extractor processes both input domains, extracting joint frequency–periodicity features to capture harmonically relevant structures. The extracted features are then fused via a Hadamard product (element-wise multiplication), aligning spectral peaks with their periodic counterparts. Finally, a fully connected (FC) layer further processes the fused features before contributing to the transcription output. As a modular neural component, the PF2N is integrated seamlessly into diverse architectures like YMT3 and YPTF without requiring structural modifications.

An overview of the PF2N's architecture is presented in Section 3.1, detailing its shared convolutional extractor in Section 3.2, feature fusion mechanism in Section 3.3, and adaptability to different transcription models in Section 3.4.

3.1. PF2N Overview

As shown in Figure 1, we apply Inverse Fourier Transform ($\mathcal{F}^{-1}$) to the given spectrogram $U(f)$, obtaining the corresponding cepstrogram $V(\frac{1}{f})$. These representations are then fed into the PF2N as follows:

$$O_{PF2N} = PF2N(U(f), V(\frac{1}{f})) \tag{1}$$

$$U_{new}(f) = U(f) + O_{PF2N} \tag{2}$$

To ensure modularity, we introduce a residual connection that combines the model's output O_{PF2N} with the original spectrogram representation $U(f)$. The residual connection ensures that the original spectral representation remains intact, enabling the PF2N output O_{PF2N} to complement and refine harmonic features by incorporating periodicity-domain enhancements, thus stabilizing training and preventing the loss of essential frequency-domain information.

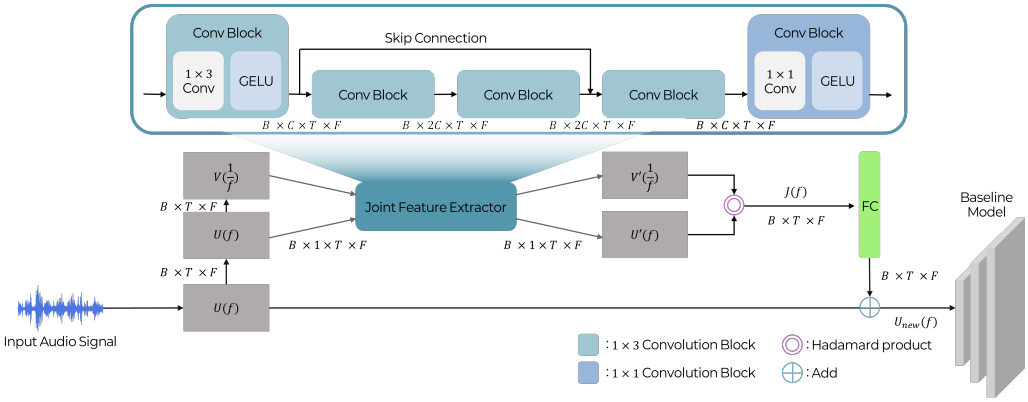

Figure 1. Overview of the Periodicity–Frequency Fusion Network (PF2N) architecture.

3.2. Joint Feature Extractor

The PF2N employs a shared convolutional module to effectively capture joint features from the spectrogram and cepstrogram. By applying identical convolutional blocks across these two inputs, the PF2N detects harmonically correlated activations inherent in both domains, enhancing fundamental frequency tracking and mitigating interference from overlapping harmonics. The joint feature extractor consists of multiple convolutional blocks (*ConvBlocks*), each extracting harmonic patterns from both representations. The feature extraction process follows

$$X^{i+1}_{c=c_{i+1}} = ConvBlock_{1\times3}(X^{i}_{c=c_i}) \tag{3}$$

$$X^{4}_{c=c_4} = ConvBlock_{1\times3}(Concat(X^{3}_{c=c_3}, X^{1}_{c=c_1})) \tag{4}$$

$$X^{5}_{c=1} = ConvBlock_{1\times1}(X^{4}_{c=c_4}) \tag{5}$$

$$= GELU(Conv_{1\times1}(X^{4}_{c=c_4})) \tag{6}$$

For $i = 1, 2, 3$, each Conv Block applies a 1×3 convolution, followed by GELU activation. The use of 1D convolutions maintains resolution along the time axis, enabling a precise temporal analysis without the loss of fidelity. In the intermediate step $X^4_{c=c_4}$, multi-level features are concatenated, satisfying the condition $c_4 = c_1 + c_3$. This concatenation integrates lower-level and higher-level convolutional features, enhancing the model's capability to represent multi-scale harmonic structures robustly. Finally, a 1×1 convolution reduces the channel dimension from c_4 to 1, and the GELU activation refines the representation before passing it to the fusion module.

3.3. Feature Fusion

After extracting features from the spectrogram $U(f)$ and cepstrogram $V(\frac{1}{f})$, we denote the processed representations as $U'(f)$ and $V'(\frac{1}{f})$. Inspired by conventional CFP methods, the PF2N fuses these representations using a Hadamard product, which emphasizes co-activated regions and reinforces harmonic alignment. This multiplication introduces nonlinear interactions: when both frequency and periodicity features are co-activated at a given location, their product yields a high response, highlighting harmonically consistent patterns; when only one is activated, the response is attenuated; and when neither is activated, the output yields a low response, effectively suppressing uncorrelated noise. The fusion process is formally expressed as follows:

$$J' = V'(1/f) \odot U'(f) \tag{7}$$

$$O_{PF2N} = FC(J') \tag{8}$$

The fused representation J' is then passed through an FC layer, ensuring the further refinement and integration of joint spectral and periodicity features before contributing to the final transcription output.

3.4. Baseline Differences and PF2N Adaptability

The YMT3 and YPTF models differ primarily in their encoder architecture and input processing pipeline. YMT3 follows the T5 architecture with sinusoidal positional encoding and directly processes a mel spectrogram representation. In contrast, YPTF replaces the standard transformer encoder with Perceiver TF. Additionally, YMT3 utilizes a single-sequence T5 decoder, generating a unified token sequence for all instruments, while YPTF employs a multi-channel T5 decoder, which assigns separate output channels to different instrument groups for structured transcription.

By integrating the PF2N into both YMT3 and YPTF, we demonstrate its adaptability across diverse model architectures. The PF2N adjusts to variations in input spectrogram resolution and encoder structures, maintaining its core frequency–periodicity fusion capability without architectural modifications.

4. Experiment

4.1. Settings

4.1.1. Dataset

We conducted our experiments using three publicly available datasets: Slakh2100 [29] and MusicNet [30], used for multi-instrument music, and MAESTRO [31], used for single-instrument music. Slakh2100 contains synthetic recordings rendered from MIDI, with a wide range of 34 instrument classes, including strings, winds, and percussion. MusicNet consists of freely licensed classical recordings featuring piano, string, and wind instruments. In contrast, MAESTRO comprises MIDI-synchronized solo piano performances with high-quality temporal alignment. Following prior work [24], we adopted an identical dataset split for training, validation, and testing to ensure comparability and consistency across model evaluations.

4.1.2. Preprocessing

To ensure consistency, all audio recordings were downsampled to 16 kHz. We applied Short-Time Fourier Transform (STFT) using a Hann window with a 2048 window size and constant padding. YMT3 uses a 512-bin mel spectrogram, while YPTF employs a 1024-bin log spectrogram.

4.1.3. Evaluation Metrics

We evaluated transcription performance using *Frame* F1, *Onset* F1, and *Onset* + *Offset* F1, following [24]. These metrics assess the alignment between predicted and ground-truth notes with respect to pitch, onset, and offset.

(1) *Frame* F1 : Let P_f and G_f denote the binary pitch activation vectors of predicted and ground-truth pitches at frame f, respectively, and let N_f be the total number of frames. *Frame* F1 is computed as follows:

$$\text{Precision} = \frac{\sum_{f=1}^{N_f} |P_f \cap G_f|}{\sum_{f=1}^{N_f} |P_f| + \epsilon} \tag{9}$$

$$\text{Recall} = \frac{\sum_{f=1}^{N_f} |P_f \cap G_f|}{\sum_{f=1}^{N_f} |G_f| + \epsilon} \tag{10}$$

$$\textit{Frame}\ \text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall} + \epsilon} \tag{11}$$

where ϵ is a small constant to avoid division by zero.

(2) *Onset* F1: A predicted note is counted as correct if it matches the pitch and if the onset is within ± 50 ms of the reference. Let $N_{\text{TP}}^{\text{onset}}$, $N_{\text{FP}}^{\text{onset}}$, and $N_{\text{FN}}^{\text{onset}}$ denote the number of true positives, false positives, and false negatives, respectively. Then, *Onset* F1 is computed as follows:

$$\textit{Onset}\ \text{F1} = 2 \cdot \frac{N_{\text{TP}}^{\text{onset}}}{2N_{\text{TP}}^{\text{onset}} + N_{\text{FP}}^{\text{onset}} + N_{\text{FN}}^{\text{onset}}} \tag{12}$$

(3) *Onset* + *Offset* F1: This stricter metric additionally requires the offset to match within $\max(50\ \text{ms}, 0.2 \cdot \text{duration})$, in addition to the onset and pitch. It is computed in the same way as *Onset* F1, using the corresponding true positives, false positives, and false negatives that satisfy both onset and offset conditions. For multi-instrument settings, we apply the same evaluation criteria but additionally require the predicted instrument label to match the ground-truth when computing true positives.

4.2. Main Results

Table 1 highlights the performance improvements achieved by integrating the PF2N into the baseline models YMT3 and YPTF. Across the Slakh2100, MusicNet, and MAESTRO datasets, the PF2N consistently enhances transcription accuracy. In YMT3, the PF2N improves *Onset* F1 by 3.3% on Slakh2100, 2.1% on MusicNet, and 0.7% on MAESTRO. For *Onset* + *Offset* F1, it achieves gains of 3.8%, 1.8%, and 3.7% on the respective datasets. While absolute improvements are larger in YMT3, the gains observed in YPTF remain meaningful given its stronger baseline and the inherently limited headroom for further improvement.

Beyond baseline differences, architectural distinctions between the two models may also contribute to the variation in the PF2N's impact. YMT3 directly processes the spectrogram with a T5-based encoder, while YPTF incorporates a pre-encoder that extracts and enhances spectral features before passing the inputs to its Perceiver TF-based encoder. Since the pre-encoder of YPTF refines the spectral representations before reaching the main encoder, the additional improvements of the PF2N may be less pronounced than those of YMT3.

Table 1. Main results (* indicates reproduced results).

Model	Slakh2100	MusicNet	MAESTRO
Frame F1 (%)			
Melodyne	47.0	13.0	41.0
ReconVAT [18]	-	48.0	-
MT3 [24]	78.0	60.0	88.0
YMT3 * [25]	66.0	57.5	82.7
YMT3 w/ PF2N	66.4	58.8	83.6
	+0.4	**+1.3**	**+0.9**
YPTF * [27]	82.8	63.7	90.6
YPTF w/ PF2N	83.4	64.5	90.8
	+0.6	**+0.8**	**+0.2**
Onset F1 (%)			
Melodyne	30.0	4.0	52.0
ReconVAT [18]	-	29.0	-
Jointist [23]	58.4	-	-
MT3 [24]	76.0	39.0	96.0
Perceiver TF [26]	80.8	-	96.7
YMT3 * [25]	53.8	37.5	88.1
YMT3 w/ PF2N	57.1	39.6	88.8
	+3.3	**+2.1**	**+0.7**
YPTF * [27]	80.9	46.1	96.9
YPTF w/PF2N	81.6	47.5	96.9
	+0.7	**+1.4**	**+0.0**
Onset + Offset F1 (%)			
Melodyne	10.0	1.0	6.0
ReconVAT [18]	-	11.0	-
Jointist [23]	26.3	-	-
MT3 [24]	57.0	21.0	84.0
YMT3 * [25]	33.4	21.2	69.1
YMT3 w/ PF2N	37.2	23.0	72.8
	+3.8	**+1.8**	**+3.7**
YPTF * [27]	65.6	29.0	87.4
YPTF w/ PF2N	66.7	30.7	87.7
	+1.1	**+1.7**	**+0.3**

4.3. Ablation Studies

Table 2 presents the instrument-wise performance improvements achieved by the PF2N in terms of *Onset* F1 on the Slakh2100 and MusicNet datasets. For the YMT3 model on Slakh2100, notable improvements are observed in instruments characterized by a rich harmonic structure, including guitar (+7.2%), string (+6.1%), organ (+7.0%), pipe (+6.6%), and reed instruments (+6.9%). In contrast, synth pad (S.pad) exhibits slight degradation (-0.1% in YMT3 and -0.3% in YPTF). With its gradual attack and long release, synth pad presents less distinct onset cues, making it harder to detect onsets using periodicity-based features. This serves as a representative failure case of the PF2N, highlighting its limitations in handling instruments with weakly pitched onsets.

Table 3 reinforces these findings under the stricter *Onset* + *Offset* F1 metric. Instruments with distinct pitched transients continue to benefit the most, such as guitar (+6.9%), reed (+5.9%), organ (+5.3%), and pipe instruments (+4.5%). Meanwhile, S.pad again shows minimal or negative change (+0.0% in YMT3 and -1.3% in YPTF). These consistent trends across both metrics indicate that the PF2N enhances transcription by emphasizing harmonic

and periodic patterns aligned with well-defined note boundaries while offering limited benefits for instruments with smooth temporal characteristics.

Table 2. Instrument-wise performance improvement on Onset F1 (%). (Bold values denote the performance differences from the corresponding baseline models.)

		Slakh2100												**MusicNet**		
Model	#Param	Piano	Bass	Drums	Guitar	Strings	Brass	Organ	Pipe	Reed	S.lead	S.pad	C.perc.	Piano	Strings	Winds
YMT3	44.9 M	51.6	64.6	62.3	45.8	34.8	22.3	19.4	26.1	28.7	25.4	16.0	18.5	40.7	28.5	30.4
YMT3 w/ PF2N	45.2 M	57.6	65.4	64.0	53.0	40.9	24.8	26.4	32.7	35.6	30.9	15.9	26.9	44.4	30.8	32.7
	+0.3 M	**+6.0**	**+0.8**	**+1.7**	**+7.2**	**+6.1**	**+2.5**	**+7.0**	**+6.6**	**+6.9**	**+5.5**	−0.1	**+8.4**	**+3.7**	**+2.3**	**+2.3**
YPTF	96.4 M	83.9	92.2	84.0	78.3	70.1	69.2	66.8	67.9	75.3	79.4	40.4	66.1	53.3	40.7	46.0
YPTF w/ PF2N	97.5 M	84.9	92.5	83.5	79.5	70.5	73.7	67.7	67.8	77.2	81.1	40.1	66.3	54.5	41.6	49.0
	+1.1 M	**+1.0**	**+0.3**	**+0.5**	**+0.8**	**+0.4**	**+4.5**	**+0.9**	**+0.1**	**+1.9**	**+1.7**	−0.3	**+0.2**	**+1.2**	**+0.9**	**+3.0**

Table 3. Instrument-wise performance improvement on Onset + Offset F1 (%). (Bold values denote the performance differences from the corresponding baseline models.)

		Slakh2100												**MusicNet**		
Model	#Param	Piano	Bass	Drums	Guitar	Strings	Brass	Organ	Pipe	Reed	S.lead	S.pad	C.perc.	Piano	Strings	Winds
YMT3	44.9 M	27.2	48.3	-	28.8	20.7	14.4	10.8	16.3	18.3	15.9	0.1	0.1	23.2	15.8	15.8
YMT3 w/ PF2N	45.2 M	32.0	49.6	-	35.7	24.1	16.5	16.1	20.8	24.2	21.4	0.1	0.1	25.5	17.5	18.4
	+0.3 M	**+4.8**	**+1.3**	-	**+6.9**	**+3.4**	**+2.1**	**+5.3**	**+4.5**	**+5.9**	**+5.5**	**+0.0**	**+0.0**	**+2.3**	**+1.7**	**+2.6**
YPTF	96.4 M	61.8	83.8	-	63.6	55.1	62.3	55.4	57.2	65.9	69.0	27.1	36.4	32.6	25.6	32.3
YPTF w/ PF2N	97.5 M	63.0	84.3	-	65.0	56.0	63.9	56.5	56.6	68.2	72.1	25.8	36.3	34.0	26.5	36.3
	+1.1 M	**+1.2**	**+0.5**	-	**+1.4**	**+0.9**	**+1.6**	**+1.1**	−0.6	**+2.3**	**+3.1**	−1.3	−0.1	**+1.4**	**+0.9**	**+4.0**

Table 4 presents an ablation analysis of the individual contributions of frequency and periodicity features. Compared to the baseline, incorporating only frequency features into the PF2N results in no change in *Frame* F1, while periodicity alone yields a gain of +0.2%. The full fusion of both domains achieves a +0.8% improvement, confirming that the PF2N effectively captures complementary harmonic information. A consistent pattern is also observed in *Onset* + *Offset* F1. Frequency and periodicity individually offer moderate gains of 0.2% and 0.5% over the baseline, and their combination results in the best performance, with a 1.7% increase. These results support the PF2N's neural design, which learns to align and integrate both domain features for robust pitch and onset recognition.

Table 4. Impacts of periodicity and frequency features in the PF2N (baseline: YPTF; dataset: MusicNet).

Model	**Frame F1 (%)**	**Onset F1 (%)**	**Onset + Offset F1 (%)**
P + F	64.5	47.5	30.7
P only	63.9	46.2	29.5
F only	63.7	46.0	29.2
None	63.7	46.1	29.0

5. Conclusions

In this paper, we introduced the PF2N, a lightweight modular component that enhances transcription performance by fusing frequency-domain and periodicity-domain features. The PF2N generalizes the classical CFP approach through a neural architecture that learns to extract and align harmonic patterns in a data-driven manner. Experiments across multiple datasets and model backbones validate its effectiveness and adaptability. Our work highlights the importance of multi-domain feature integration for improving structured audio representation and recognition. We believe that the PF2N offers a flexible tool for advancing neural pattern recognition in music, and future work will explore its application to real-time transcription, instrument identification, and source separation.

Author Contributions: Conceptualization, T.K.; methodology, T.K.; software, T.K.; validation, M.-J.K. and C.W.A.; formal analysis, T.K.; investigation, T.K.; resources, C.W.A.; data curation, T.K.; writing—original draft preparation, T.K.; writing—review and editing, M.-J.K. and C.W.A.; visualization, T.K.; supervision, M.-J.K.; project administration, C.W.A.; funding acquisition, M.-J.K. and C.W.A. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (RS-2023-00247900); the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. 2019-0-01842, Artificial Intelligence Graduate School Program (GIST) & the Artificial Intelligence Convergence Innovation Human Resources Development (RS-2023-00256629)); and the ITRC (Information Technology Research Center) Support Program (IITP-2025-RS-2024-00437718).

Data Availability Statement: The data presented in this study are openly available in [29–31].

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

References

1. Benetos, E.; Dixon, S.; Duan, Z.; Ewert, S. Automatic Music Transcription: An Overview. *IEEE Signal Process. Mag.* **2019**, *36*, 20–30. [CrossRef]
2. Huang, C.A.; Vaswani, A.; Uszkoreit, J.; Simon, I.; Hawthorne, C.; Shazeer, N.; Dai, A.M.; Hoffman, M.D.; Dinculescu, M.; Eck, D. Music Transformer: Generating Music with Long-Term Structure. In Proceedings of the 7th International Conference on Learning Representations, ICLR, New Orleans, LA, USA, 6–9 May 2019.
3. Dong, H.W.; Chen, K.; Dubnov, S.; McAuley, J.; Berg-Kirkpatrick, T. Multitrack Music Transformer. In Proceedings of the ICASSP 2023–2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, 4–10 June 2023.
4. Armentano, M.G.; De Noni, W.A.; Cardoso, H.F. Genre classification of symbolic pieces of music. *J. Intell. Inf. Syst.* **2017**, *48*, 579–599. [CrossRef]
5. Tsai, T.; Ji, K. Composer style classification of piano sheet music images using language model pretraining. In Proceedings of the 21th International Society for Music Information Retrieval Conference, ISMIR, Online, 7–12 November 2021; pp. 176–183.
6. Simonetta, F.; Chacón, C.E.C.; Ntalampiras, S.; Widmer, G. A Convolutional Approach to Melody Line Identification in Symbolic Scores. In Proceedings of the 20th International Society for Music Information Retrieval Conference, ISMIR, Delft, The Netherlands, 4–8 November 2019; pp. 924–931.
7. Rafii, Z.; Liutkus, A.; Stöter, F.R.; Mimilakis, S.I.; FitzGerald, D.; Pardo, B. An overview of lead and accompaniment separation in music. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2018**, *26*, 1307–1335. [CrossRef]
8. Poliner, G.E.; Ellis, D.P. A discriminative model for polyphonic piano transcription. *EURASIP J. Adv. Signal Process.* **2006**, *2007*, 1–9. [CrossRef]
9. Costantini, G.; Perfetti, R.; Todisco, M. Event based transcription system for polyphonic piano music. *Signal Process.* **2009**, *89*, 1798–1811. [CrossRef]
10. O'Hanlon, K.; Plumbley, M.D. Polyphonic piano transcription using non-negative Matrix Factorisation with group sparsity. In Proceedings of the 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Florence, Italy, 4–9 May 2014; pp. 3112–3116.
11. Gao, L.; Su, L.; Yang, Y.H.; Lee, T. Polyphonic piano note transcription with non-negative matrix factorization of differential spectrogram. In Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 5–9 March 2017; pp. 291–295.
12. Wu, H.; Marmoret, A.; Cohen, J.E. Semi-supervised convolutive NMF for automatic piano transcription. *arXiv* **2022**, arXiv:2202.04989.
13. Marolt, M. Transcription of polyphonic piano music with neural networks. In Proceedings of the 2000 10th Mediterranean Electrotechnical Conference, Information Technology and Electrotechnology for the Mediterranean Countries, Proceedings. MeleCon 2000 (Cat. No.00CH37099), Lemesos, Cyprus, 29–31 May 2000; Volume 2, pp. 512–515.
14. Van Herwaarden, S.; Grachten, M.; De Haas, W.B. Predicting expressive dynamics in piano performances using neural networks. In Proceedings of the 15th International Society for Music Information Retrieval Conference, ISMIR, International Society for Music Information Retrieval, Taipei, Taiwan, 27–31 October 2014; pp. 45–52.
15. Sigtia, S.; Benetos, E.; Dixon, S. An End-to-End Neural Network for Polyphonic Piano Music Transcription. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2016**, *24*, 927–939. [CrossRef]

16. Hawthorne, C.; Elsen, E.; Song, J.; Roberts, A.; Simon, I.; Raffel, C.; Engel, J.H.; 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.
17. Kong, Q.; Li, B.; Song, X.; Wan, Y.; Wang, Y. High-resolution piano transcription with pedals by regressing onset and offset times. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2021**, *29*, 3707–3717. [CrossRef]
18. Cheuk, K.W.; Herremans, D.; Su, L. ReconVAT: A Semi-Supervised Automatic Music Transcription Framework for Low-Resource Real-World Data. In Proceedings of the MM '21: ACM Multimedia Conference, Virtual Event, 20–24 October 2021; pp. 3918–3926.
19. Manilow, E.; Seetharaman, P.; Pardo, B. Simultaneous Separation and Transcription of Mixtures with Multiple Polyphonic and Percussive Instruments. In Proceedings of the 2020 IEEE International Conference on Acoustics, Speech and Signal Processing, Barcelona, Spain, 4–8 May 2020; pp. 771–775.
20. Tanaka, K.; Nakatsuka, T.; Nishikimi, R.; Yoshii, K.; Morishima, S. Multi-Instrument Music Transcription Based on Deep Spherical Clustering of Spectrograms and Pitchgrams. In Proceedings of the 21th International Society for Music Information Retrieval Conference, ISMIR, Montreal, QC, Canada, 11–16 October 2020; pp. 327–334.
21. Lin, L.; Xia, G.; Kong, Q.; Jiang, J. A unified model for zero-shot music source separation, transcription and synthesis. In Proceedings of the 22nd International Society for Music Information Retrieval Conference, ISMIR, Online, 7–12 November 2021; pp. 381–388.
22. Wu, Y.T.; Chen, B.; Su, L. Multi-Instrument Automatic Music Transcription With Self-Attention-Based Instance Segmentation. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2020**, *28*, 2796–2809. [CrossRef]
23. Cheuk, K.W.; Choi, K.; Kong, Q.; Li, B.; Won, M.; Hung, A.; Wang, J.; Herremans, D. Jointist: Joint Learning for Multi-instrument Transcription and Its Applications. *arXiv* **2022**, arXiv:2206.10805.
24. Gardner, J.; Simon, I.; Manilow, E.; Hawthorne, C.; Engel, J.H. MT3: Multi-Task Multitrack Music Transcription. In Proceedings of the The Tenth International Conference on Learning Representations, ICLR, Virtual Event, 25–29 April 2022.
25. Chang, S.; Dixon, S.; Benetos, E. YourMT3: A toolkit for training multi-task and multi-track music transcription model for everyone. In Proceedings of the Digital Music Research Network One-day Workshop (DMRN+ 17), London, UK, 20 December 2022.
26. Lu, W.T.; Wang, J.; Hung, Y. Multitrack Music Transcription with a Time-Frequency Perceiver. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP, Rhodes Island, Greece, 4–10 June 2023; pp. 1–5.
27. Chang, S.; Benetos, E.; Kirchhoff, H.; Dixon, S. YourMT3+: Multi-Instrument Music Transcription with Enhanced Transformer Architectures and Cross-Dataset STEM Augmentation. In Proceedings of the 34th IEEE International Workshop on Machine Learning for Signal Processing, MLSP 2024, London, UK, 22–25 September 2024; pp. 1–6.
28. Su, L.; Yang, Y.H. Combining Spectral and Temporal Representations for Multipitch Estimation of Polyphonic Music. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2015**, *23*, 1600–1612. [CrossRef]
29. Manilow, E.; Wichern, G.; Seetharaman, P.; Le Roux, J. Cutting Music Source Separation Some Slakh: A Dataset to Study the Impact of Training Data Quality and Quantity. In Proceedings of the Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), New Paltz, NY, USA, 20–23 October 2019.
30. Thickstun, J.; Harchaoui, Z.; Kakade, S.M. Learning Features of Music From Scratch. In Proceedings of the 5th International Conference on Learning Representations, ICLR, Toulon, France, 24–26 April 2017.
31. Hawthorne, C.; Stasyuk, A.; Roberts, A.; Simon, I.; Huang, C.A.; Dieleman, S.; Elsen, E.; Engel, J.H.; Eck, D. Enabling Factorized Piano Music Modeling and Generation with the MAESTRO Dataset. In Proceedings of the 7th International Conference on Learning Representations, ICLR, New Orleans, LA, USA, 6–9 May 2019.
32. Su, L.; Chuang, T.; Yang, Y. Exploiting Frequency, Periodicity and Harmonicity Using Advanced Time-Frequency Concentration Techniques for Multipitch Estimation of Choir and Symphony. In Proceedings of the 17th International Society for Music Information Retrieval Conference, ISMIR, New York, NY, USA, 7–11 August 2016; pp. 393–399.
33. Su, L. Vocal Melody Extraction Using Patch-Based CNN. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, Calgary, AB, Canada, 15–20 April 2018; pp. 371–375.
34. Wu, Y.T.; Chen, B.; Su, L. Automatic Music Transcription Leveraging Generalized Cepstral Features and Deep Learning. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 401–405.
35. Matsunaga, T.; Saito, H. Multi-Layer Combined Frequency and Periodicity Representations for Multi-Pitch Estimation of Multi-Instrument Music. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2024**, *32*, 3171–3184. [CrossRef]