symmetry

MDPI

Article

Dual Attention Network for Pitch Estimation of Monophonic Music

Wenfang Ma [1,2], Ying Hu [1,2,*] and Hao Huang [1,3]

[1] School of Information Science and Engineering, Xinjiang University, Urumqi 830046, China; mawenfang@stu.xju.edu.cn (W.M.); huanghao@xju.edu.cn (H.H.)
[2] Key Laboratory of Signal Detection and Processing in Xinjiang, Urumqi 830046, China
[3] Key Laboratory of Multilingual Information Technology in Xinjiang, Urumqi 830046, China
* Correspondence: huying@xju.edu.cn; Tel.: +86-186-0991-4062

Abstract: The task of pitch estimation is an essential step in many audio signal processing applications. In this paper, we propose a data-driven pitch estimation network, the Dual Attention Network (DA-Net), which processes directly on the time-domain samples of monophonic music. DA-Net includes six Dual Attention Modules (DA-Modules), and each of them includes two kinds of attention: element-wise and channel-wise attention. DA-Net is to perform element attention and channel attention operations on convolution features, which reflects the idea of "symmetry". DA-Modules can model the semantic interdependencies between element-wise and channel-wise features. In the DA-Module, the element-wise attention mechanism is realized by a Convolutional Gated Linear Unit (ConvGLU), and the channel-wise attention mechanism is realized by a Squeeze-and-Excitation (SE) block. We explored three kinds of combination modes (serial mode, parallel mode, and tightly coupled mode) of the element-wise attention and channel-wise attention. Element-wise attention selectively emphasizes useful features by re-weighting the features at all positions. Channel-wise attention can learn to use global information to selectively emphasize the informative feature maps and suppress the less useful ones. Therefore, DA-Net adaptively integrates the local features with their global dependencies. The outputs of DA-Net are fed into a fully connected layer to generate a 360-dimensional vector corresponding to 360 pitches. We trained the proposed network on the iKala and MDB-stem-synth datasets, respectively. According to the experimental results, our proposed dual attention network with tightly coupled mode achieved the best performance.

Keywords: pitch estimation; dual attention; element-wise attention; channel-wise attention

Citation: Ma, W.; Hu, Y.; Huang, H. Dual Attention Network for Pitch Estimation of Monophonic Music. *Symmetry* **2021**, *13*, 1296. https://doi.org/10.3390/sym13071296

Academic Editor: Rafał Zdunek

Received: 15 April 2021
Accepted: 15 July 2021
Published: 19 July 2021

1. Introduction

F0, or pitch, is one of the most useful acoustical features that determines an audible pitch level. Pitch estimation is important in monophonic or polyphonic music signal processing. The monophonic pitch tracking method is used to generate pitch labels for multi-track datasets [1] or as a core step of melody extraction algorithms [2,3].

In recent years, this research has attracted increasing attention with the demand for singing processing [4], music information retrieval [5], large-scale analysis of different musical styles [6], and the automatic transcription of music [7]. F0 is the lowest frequency of quasi-periodic vibration of the vocal cords. Pitch is a perceptual property, and F0 is a physical property of audio. However, the pitch is determined by the F0. Despite this important distinction, pitch and F0 are generally used interchangeably outside the field of psychoacoustics.

Traditionally, simple signal processing approaches have been proposed to F0 estimators, working either in the time domain, in the frequency domain, or both, followed by a post-processing algorithm to smooth the pitch trajectory. Time-domain F0 estimation methods include YIN [8], which is based on the auto-correlation function. The pYIN [9] algorithm is an improved method of the YIN and adopts an HMM in post-processing to

improve the robustness. Frequency domain methods utilize template matching with the spectrum of a sawtooth waveform (e.g., SWIPE [10]).

YAAPT [11] combines frequency domain processing, time-domain processing, and normalized cross-correlation. Nebula [12] uses F0 and voicing status estimation algorithms for high-quality speech analysis/synthesis. These are both time-frequency domain F0 estimation methods. A obvious trend in the methods [8–12] is that the derivation of a good F0 estimation method depends on devising a robust candidate generating function or sophisticated pre- or post-processing steps. These are all not learned from data directly. Even the best performing algorithms, such as the pYIN [9] method can also produce noisy results for challenging audio recordings.

In recent years, neural network pitch estimators of audio have emerged. For example, Zhang et al. [13] used the PEFAC algorithm [14] to extract spectral domain features from each frame as input and proposed to exploit RNN-BLSTM to model the two pitch contours of a mixture of two speech signals. However, recently, some data-driven methods using various neural networks have been proposed for monophonic pitch estimation [15,16] and multi-pitch [17,18] or melody tracking [19–23], outperforming previous results.

In [24], an end-to-end regression model based on a neural network was proposed, where a voice detector and a proposed F0 value estimator work jointly to highlight the trajectory of the pitch. Manu Airaksinen et al. [25] explored multiple speech data augmentation methods for the task of pitch estimation with neural networks under noisy conditions. CREPE [26] is a deep convolutional neural network for pitch estimation that performs directly on the time-domain waveform and produced outstanding results.

Luc Ardaillon et al. [27] proposed a fully-convolutional network (FCN) architecture for pitch estimation of speech that could reduce the computational load making it more suitable for real-time purposes. Dong et al. [28] proposed a deep convolution residual network for vocal pitch extraction in polyphonic music. Gfeller et al. [29] proposed the SPICE model, a self-supervised pitch estimation algorithm for monophonic audio.

CREPE [26] produces outstanding results when using a neural network to estimate the pitch. The input of CREPE is 1024 samples (frame length) extracted from a waveform. Six Convolutional neural network (CNN) layers produce a 2048-dimensional feature representation, which is then fed into a fully connected layer to obtain a 360-dimensional output vector $\hat{\mathbf{y}}$.

Dauphin et al. [30] first proposed a gating mechanism for a language model in 2017. Since then, the gating mechanism, also known as Gated Linear Units (GLUs), has been widely used in speech audio processing. Tang and Wang [31] proposed a convolutional recurrent neural network and incorporated gated linear units for complex spectral mapping, which amounts to a causal system for monaural speech enhancement. CNN incorporating a gating mechanism has been proposed for speech enhancement [32], speech separation [33], and singing voice and accompaniment separation [34].

This CNN incorporates a gating mechanism denoted as convolutional gated linear unit (ConvGLU), which can re-weight each feature element in feature maps. The element-wise attention mechanism in this paper is realized by using ConvGLU, which will be discussed in more detail later. Squeeze-and-Excitation Networks (SENets) were first proposed by Hu et al. [35]. The SE block adaptively recalibrates channel-wise feature maps by explicitly modeling the interdependencies between channels.

Since then, SE block, also known as a Channel-wise Attention mechanism, has been widely applied to the image processing field such as Densely Squeeze-and-Excitation Network (DSENet) [36], Deep Residual Squeeze-and-Excitation Network (DRSEN) [37], and SERAN [38], which integrate Squeeze-and-Excitation (SE) modules and attention modules into a ResNet-based classifier. Wu et al. [39] introduced a SENet into a CNN. The channel-wise attention mechanism in this paper is realized by a Squeeze-and-Excitation block, which will be discussed in Section 2.3.

Motivated by Gated Linear Units (GLUs) [30] and SENets [35], we propose a data-driven Dual Attention Network (DA-Net) for pitch estimation. DA-Net is to perform

element attention and channel attention operations on convolution features, which reflects the idea of "symmetry". DA-Net adopts the CREPE structure and applies dual attention modules to extract feature maps instead of conventional CNN in CREPE. The outputs of DA-Net are fed into the post-processing part to obtain the final F0. Fu Jun et al. [40] initially proposed dual attention, which process the task of scenario segmentation by capturing rich context dependencies through the self-attention mechanism.

It consists of two attention modules, channel attention and spatial attention. The spatial attention module selectively aggregate the features of each location through the weighted sum of all location features and it aims to learn relationships between pixels in the pictures, which actually is the implementation of non-local idea. Yu, S. et al [41] proposed frequency-temporal attention, which includes frequency and temporal attention. Frequency attention aims to select the stimulated frequency band in the spectrum, just as in the cochlea, while the purpose of time attention is to simulate the auto-correlation in the cortex. Zhibin Hu et al. [42] proposed a dual attention network for image and text matching, which can infer the importance of all the words to each image region and infer the importance of all the image regions to each word.

Bo Li [43] proposed a graph-enhanced double-attention network (GEDA) for document-level relational extraction, which can better describe the complex interaction between the sentences and potential relational instances, thus, improving the reasoning ability between the sentences. Yinglin Zhu et al. [44] proposed an interactive dual attention model to interactively learn the representation between the contextual semantics and affective orientation information. Jie Wan et al. [45] designed a dual attention module to help networks capture spatially and channel dependent road features for better semantic inference of occluded roads.

Dual attention network in this paper includes element-wise attention realized by ConvGLU and channel-wise attention realized by SE block. The main contributions are as follows:

(i) For pitch estimation of monophonic music, we propose a data-driven DA-Net integrating the element-wise attention mechanism and the channel-wise attention mechanism.
(ii) We explored three combination modes of the two attention mechanisms: serial mode, parallel mode, and tightly coupled mode. According to the experiments, the Dual Attention network with Tightly Coupled mode (DA-TC) obtained the best results.
(iii) We validated our network on the iKala and MDB-stem-synth datasets, respectively. The DA-TC achieved improvement comparing with CREPE, especially on MDB-stem-synth.

The rest of the paper is organized as follows. Section 2 describes the proposed model. Section 3 details the experimental setup, and Section 4 shows the experimental results. We draw our conclusions in Section 5.

2. Proposed Model

We first introduce the proposed DA-Net for pitch estimation of monophonic music. Then, we explore three kinds of combination modes of two attention mechanisms and describe the implementation of the tightly coupled mode in detail. Finally, we introduce two kinds of attention mechanisms.

2.1. Dual Attention Network

Figure 1 shows the structure of the proposed DA-Net for pitch estimation of monophonic music. The DA-Net including six DA-Modules processes the time-domain waveform directly to obtain the estimated pitch. Each DA-Module includes an element-wise attention mechanism and channel-wise attention mechanism. The input is 1024 samples (frame length). After being normalized to a zero mean and unit variance, the inputs undergo six DA-Modules to obtain a potential representation of 2048 dimensions.

Finally, DA-Net outputs a 360-dimensional vector through a fully connected layer. In summary, this network performs a frame-by-frame classification task of pitch estimation.

As seen in Figure 1, DA-Net is an end-to-end pitch estimation network. The DA-Module indicated by the pink box includes two kinds of attention mechanisms, which will be discussed in more detail in Sections 2.3 and 2.4.

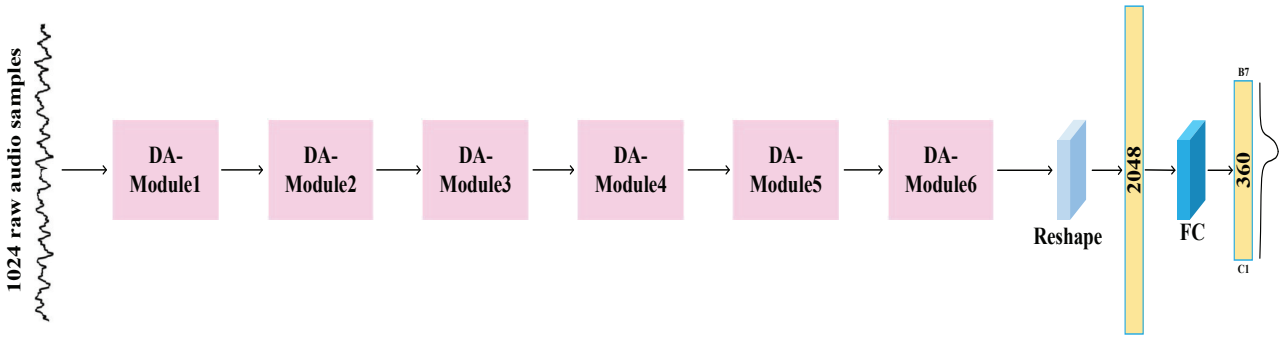

Figure 1. Proposed Dual Attention Network (DA-Net).

2.2. Dual Attention Module

The DA-Module in Figure 1 is designed to extract feature maps for pitch estimation. Each DA-Module includes an element-wise attention mechanism and channel-wise attention mechanism. Here, we explored three kinds of combination modes (serial mode, parallel mode, and tightly coupled mode) of the two kinds of attention mechanisms. Figure 2 shows a schematic diagram of the above three kinds of combination modes. Figure 2a denotes the diagram of the serial mode, where the inputs are first fed into the element-wise attention block and then the outputs are further fed into the channel-wise attention block. Channel-wise attention is realized by SE block.

The left half part of Figure 2a realizes the element-wise attention mechanism. The parallel mode (Figure 2b) includes two branches, where the upper half branch applies convolution to extract the feature maps of inputs before sending it to the channel attention block. The lower half branch of Figure 2b denotes an element-wise attention block. The serial mode and the parallel mode are both loosely coupled modes. The tightly coupled mode in Figure 2c denotes that the convolutional feature maps in the element-wise attention block are fed into channel-wise attention block. Next, we describe the tightly coupled mode in detail.

Tightly Coupled Mode

For the tightly coupled mode of two kinds of attention mechanisms in Figure 2c, a detailed structure diagram is shown in Figure 3. In Figure 3, two dashed frames denote the element-wise attention block and channel-wise attention block, respectively. In the top half part of the element-wise attention block, the feature maps after a 1-dimension CNN operation are fed into the channel-wise attention block to learn the attention among different channels, meanwhile, this serves as the feature maps of the element-wise attention block.

The element-wise attention mechanism is realized by the Convolutional Gated Linear Unit (ConvGLU), and the channel-wise attention mechanism is realized by the SE block. The outputs of the element-wise attention block and channel-wise attention block are summed and followed with a BN and a max pooling operation. Thus, two kinds of attention blocks are tightly coupled together. As seen in the right half part in Figure 3, cubes with different colors denote conv1d, global average pooling, fully connected layer, Relu, batch normalization, and max pooling operations. Max pooling with the size of 1×2 or 1×4 are used to downsample the time dimension. Our experimental results show that the tightly coupled mode outperformed the other two modes.

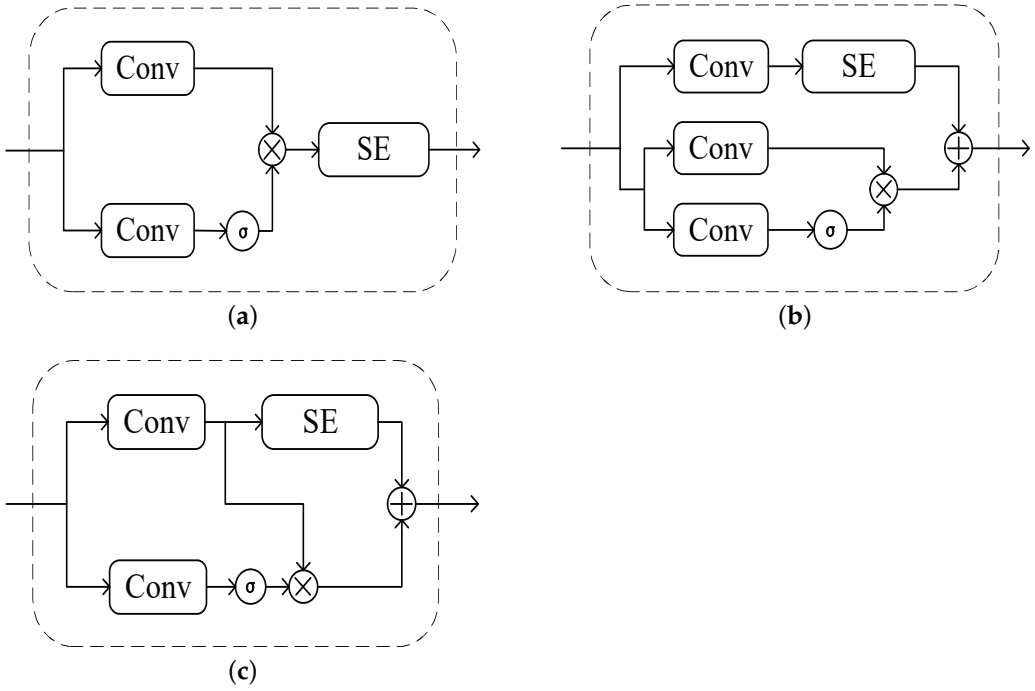

Figure 2. Schematic diagram of the combination modes of two kinds of attention mechanisms. (**a**) The serial mode. (**b**) The parallel mode. (**c**) The tightly coupled mode. σ denotes the sigmoid activation function.

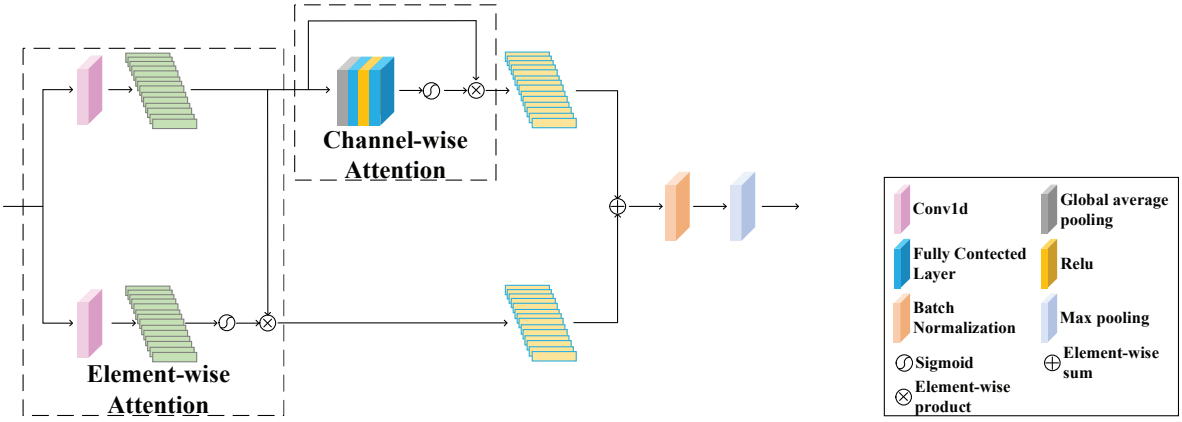

Figure 3. Illustration of the tightly coupled mode. The dashed blocks in the left half part denote the channel-wise attention block and the element-wise attention block. Cubes denote conv1d, global average pooling, fully connected layer, Relu, batch normalization, and max pooling.

For the two different CNNs in ConvGLU, the kernel size, the stride, the filter size, and the pooling size are all set with the same number. The CNN in the lower half part of ConvGLU learns the importance of each element in the feature maps and then recalibrates element-wise features by the multiplication operation. The channel-wise attention mechanism adaptively recalibrates channel-wise feature responses by explicitly modeling interdependencies between channels, selectively emphasizing informative features and suppressing less useful ones. The parameters used for ConvGLUs of the DA-Net structure are listed in Table 1.

Table 1. Details of the parameters used for ConvGLUs of the DA-Net structure.

Name	Kernel Size/Stride/Pooling Size	Output Size (Channel, Time)
Input	-	(1, 1024)
DA-Module1	512/4/2	(1024, 128)
DA-Module2	64/1/2	(128, 64)
DA-Module3	64/1/2	(128, 32)
DA-Module4	64/1/2	(128, 16)
DA-Module5	64/1/2	(256, 8)
DA-Module6	64/1/2	(512, 4)
FC layer	-	(1, 360)

$\mathbf{X}_0$ is the input of the DA-Net. $\mathbf{X}_i, i \in \{1,2,3,4,5,6\}$ are the outputs of each module in the network, where i is the index of the downsampling modules. Each DA-Module includes the outputs of two attention blocks, $\mathbf{X}_{i,0}$ and $\mathbf{X}_{i,1}$.

$\mathbf{X}_{i,0}$ is the output of the Channel-wise Attention block.

$$\mathbf{X}_{i,0} = \mathbf{F}_{SE}(\mathbf{W}_i * \mathbf{X}_{i-1}), \tag{1}$$

where $\mathbf{W}_i$ is the weight matrix of the upper half part CNN of the ConvGLU, $\mathbf{X}_{i-1}$ denotes the output of the former, and $*$ denotes a convolution operation. The function $\mathbf{F}_{SE}(\cdot)$ denotes the Squeeze-and-Excitation operating.

$\mathbf{X}_{i,1}$ is the output of the Element-wise Attention block.

$$\mathbf{X}_{i,1} = \mathbf{H}_{GLU}(\mathbf{X}_{i-1}) = (\mathbf{W}_i * \mathbf{X}_{i-1}) \odot \sigma(\mathbf{W}'_i * \mathbf{X}_{i-1}), \tag{2}$$

where $\mathbf{W}'_i$ is the weight matrix of the lower half part CNN of the ConvGLU, σ is the sigmoid activation function, and $\odot$ denotes the element-wise product. $\mathbf{H}_{GLU}(\cdot)$ is the Convolutional Gated Linear Unit (ConvGLU) operating.

$$\mathbf{X}_i = \mathbf{X}_{i,0} + \mathbf{X}_{i,1} \tag{3}$$

where $i \in \{1,2,3,4,5,6\}$, $\mathbf{X}_0$ is the input of the DA-Net and $\mathbf{X}_i$ are the outputs of each module in the DA-Net.

2.3. Channel-Wise Attention Mechanism

The channel-wise attention mechanism is realized by the SE block [35], which focuses on the relationships among the channels of feature maps. The diagram of the channel-wise attention mechanism is shown in Figure 4.

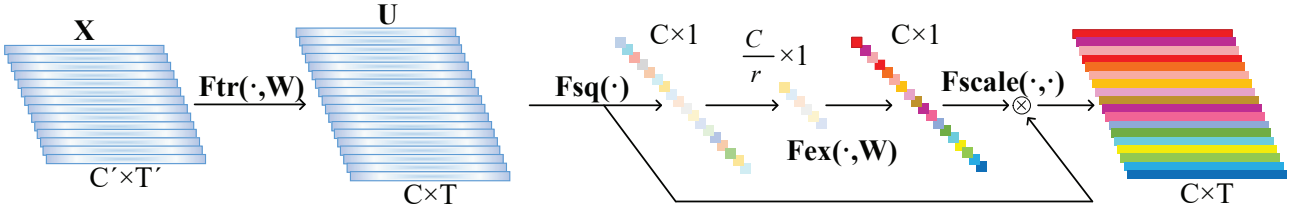

Figure 4. Channel-wise Attention Mechanism.

The SE block is a computational unit that can be constructed for any given transformation $\mathbf{F}_{tr} : \mathbf{X} \to \mathbf{U}, \mathbf{X} \in \mathbb{R}^{C' \times T'}, \mathbf{U} \in \mathbb{R}^{C \times T}$. $\mathbf{F}_{tr}$ here refers to the convolutional operator.

$$\mathbf{U} = \mathbf{F}_{tr}(\mathbf{X}) = \mathbf{W} * \mathbf{X} = [\mathbf{w}_1 * \mathbf{X}, \mathbf{w}_2 * \mathbf{X}, \ldots, \mathbf{w}_C * \mathbf{X}] \tag{4}$$

$\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_C]$ denotes the learnable set of filter kernels. $\mathbf{U} = [\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_C]$ denotes the output of the $\mathbf{F}_{tr}$ operation.

Squeeze: Global Information Embedding

First, the global spatial information is squeezed into a channel descriptor. This is realized by using global average pooling to obtain channel-wise statistics. A statistic $\mathbf{z} \in \mathbb{R}^C$ is generated by compressing $\mathbf{U}$ through $1 \times T$, where the c-th element of $\mathbf{z}$ is calculated by:

$$z_c = \mathbf{F}_{sq}(\mathbf{u}_c) = \frac{1}{T}\sum_{t=1}^{T} u_c(t) \tag{5}$$

Excitation: Adaptive Recalibration

To make use of the information aggregated in the squeeze operation, an excitation operation is followed with aiming to capture the channel-wise dependencies. A simple gating mechanism is employed with a sigmoid activation to realize an excitation operation.

$$\mathbf{s} = \mathbf{F}_{ex}(\mathbf{z}, \mathbf{W}) = \sigma(\mathbf{W}_2\, \delta(\mathbf{W}_1 \mathbf{z})) \tag{6}$$

where δ refers to the ReLU activation function,$\mathbf{W}_1 \in \mathbb{R}^{\frac{C}{r} \times C}$ and $\mathbf{W}_2 \in \mathbb{R}^{C \times \frac{C}{r}}$. We parameterize the gating mechanism by forming a bottleneck with two fully connected (FC) layers. There exists a dimensionality-reduction layer with the parameters $\mathbf{W}_1$ and the reduction ratio of r, a ReLU activation function and then a dimensionality-increasing layer with parameters $\mathbf{W}_2$. The reduction ratio r is an important hyperparameter that allows reducing the capacity and computational cost of the SE blocks in the model. r was set with 16 in our experiments. The final output of the SE block is obtained by rescaling the convolution output $\mathbf{U}$:

$$\tilde{\mathbf{x}}_c = \mathbf{F}_{scale}(s_c, \mathbf{u}_c) = s_c \odot \mathbf{u}_c, \tag{7}$$

where $\tilde{\mathbf{X}} = [\tilde{\mathbf{x}}_1, \tilde{\mathbf{x}}_2, \ldots, \tilde{\mathbf{x}}_C]$ is the output of the SE block, and $\mathbf{F}_{scale}(s_c, \mathbf{u}_c)$ refers to the channel-wise multiplication between the feature map $\mathbf{u}_c$ and the scalar s_c. The network with embedded SE block can increase its sensitivity to informative features so that they can be exploited by subsequent transforms and can suppress less useful ones.

2.4. Element-Wise Attention Mechanism

The element-wise attention mechanism is realized by ConvGLUs. ConvGLUs control the flow of information across the network, which may allow more complex interactions to be modeled [31]. Dauphin et al. [30] introduced the gating mechanism as follows:

$$\mathbf{Y}_1 \odot \sigma(\mathbf{Y}_2) = (\mathbf{X} * \mathbf{W}_1 + b_1) \odot \sigma(\mathbf{X} * \mathbf{W}_2 + b_2), \tag{8}$$

and the gradient of GLUs:

$$\nabla[\mathbf{Y}_1 \odot \sigma(\mathbf{Y}_2)] = \nabla\mathbf{Y}_1 \odot \sigma(\mathbf{Y}_2) + \mathbf{Y}_1 \odot \sigma'(\mathbf{Y}_2)\nabla\mathbf{Y}_2 \tag{9}$$

has a path $\nabla\mathbf{Y}_1 \odot \sigma(\mathbf{Y}_2)$ without downscaling factors for the activated gating units in $\sigma(\mathbf{Y}_2)$. This can be thought of as a multiplicative skip connection that helps gradients flow through the layers. Therefore, ConvGLU helps to reduce the vanishing gradient problem. A Convolutional Gated Linear Unit (denoted as "ConvGLU") block is illustrated in Figure 5.

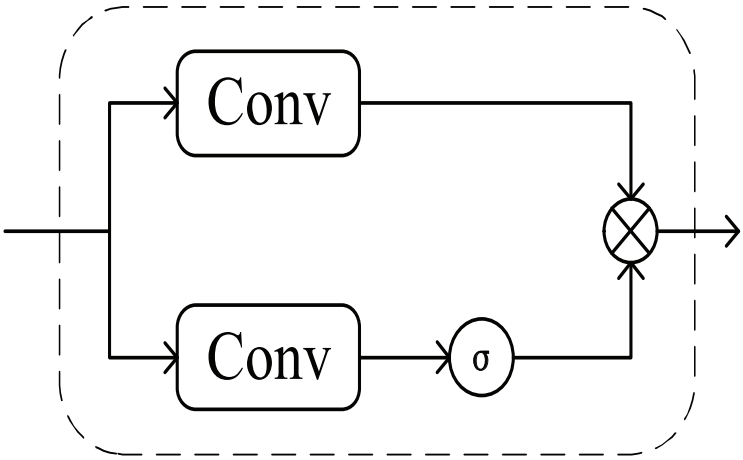

Figure 5. Diagram of the Convolutional Gated Linear Unit (ConvGLU) block.

In our proposed DA-Net, each element-wise attention block adopts ConvGLUs. As we can see from the above figure, the input of ConvGLU performs two parallel convolutional operations. The size of filters (with different weights), the number of output channels and the strides are all identical in two parallel CNNs. The output of one CNN with a sigmoid activation function is multiplied by the output of another CNN to obtain the final result of ConvGLU. The ConvGLU can be thought of as performing the element-wise attention operation.

3. Experiments

3.1. Datasets and Setting

We implemented experiments on the iKala [46] and MDB-stem-synth [47] datasets. The iKala dataset contains 352 song segments, where only 252 song clips were released as a public subset for evaluation. Each of the song segments is 30 s long with a sampling rate of 44,100 Hz. All the clips were recorded Chinese pop songs performed by professional singers. The music accompaniment and the singing voice are recorded at the left and right channels respectively. Only the singing voice was used as the samples in this paper.

The MDB-stem-synth is a collection of 230 monophonic audio clips from MedleyDB, which uses an analysis/synthesis method [47] to generate the synthesized audio with perfect ground truth F0 annotation that maintains the timbre and dynamic tracking of the original audio signal. This dataset consists of 230 tracks with 25 instruments, totaling 15.56 h of audio, hence, the name MDB-stem-synth. We performed the same processing on the iKala and MDB-stem-synth datasets.

First, the raw audio samples were downsampled to 16 kHz. The input of the pitch estimation model was 1024 samples (frame length) with the hop sizes of 160 samples in the MDB-stem-synth dataset and 512 samples in the iKala dataset. Second, before being sent to the network all samples were normalized to be with zero mean and unit variance. The network proposed in this paper performs a frame-by-frame classification task. We trained the network on the iKala and MDB-stem-synth datasets, respectively.

The networks were trained to minimize the binary cross-entropy between the predicted vector $\hat{\mathbf{y}}$ and the target vector $\mathbf{y}$:

$$\mathcal{L}(\mathbf{y}, \hat{\mathbf{y}}) = \sum_{i=1}^{360} (-y_i \log \hat{y}_i - (1 - y_i) \log(1 - \hat{y}_i)) \tag{10}$$

where both y_i and $\hat{y}_i$ are all between 0 and 1. The ADAM optimizer was adopted [48], and the initial learning rate was 10^{-4} with the decay rate of $\beta_1 = 0.96$. Each DA-Module is followed by batch normalization [49], max pooling, and a dropout layer [50] with a dropout probability of 0.25. The batch size in this method is 512.

3.2. Label Processing and Postprocessing

Similar to the CREPE described in [26], pitch estimation is regarded as a classification task. The DA-Net takes the original waveform as input and outputs the probability vector of F0 belonging to each possible output pitch class. Cent is a unit representing the intervals relative to a reference pitch f_{ref} in Hz, defined as follows:

$$c = 1200 \cdot \log_2 \frac{f}{f_{ref}} \tag{11}$$

where $f_{ref} = 10$ Hz in this paper. Each dimension of the output vector corresponds to the frequency bin that covers a frequency range from 32.7 Hz to 1975.5 Hz with 20 cents of intervals. As in [17], to soften the penalty for approaching the correct prediction value, the ground-truth label is Gaussian-blurred so that the energy around the label frequency decays with a standard deviation of 25 cents:

$$y_i = exp\left(- \frac{(c_i - c_{true})^2}{2 \cdot 25^2} \right) \tag{12}$$

Therefore, high activation in the last fully connected layer denotes that the input audio is likely to have a pitch that is close to the associated pitches of the nodes with this high activation.

Through the predicted vector $\hat{\mathbf{y}}$, the resulting pitch estimation $\hat{\mathbf{c}}$, in cents, is computed as follows.

$$\hat{c} = \frac{\sum_{i=I-4}^{I+4} \hat{y}_i c_i}{\sum_{i=I-4}^{I+4} \hat{y}_i}, I = argmax(\hat{y}_i) \tag{13}$$

This value can then be converted from cents back to Hz to obtain the final F0.

$$\hat{f} = f_{ref} \cdot 2^{\hat{c}/1200} \tag{14}$$

3.3. Evaluation Metrics

The evaluation of the proposed network is measured at both the raw pitch accuracy (RPA) and the raw chroma accuracy (RCA), with a threshold of 50 cents. We used the reference implementation provided in mir_eval [51] to compute the evaluation metrics. A brief introduction of the evaluation metrics used in this paper is shown below:

RPA: The proportion of voiced frames where the estimated pitch is within $\pm\frac{1}{4}$ tone (50 cents) of the ground truth pitch.
RCA: The proportion of voiced frames in which the estimated pitch and the ground truth pitch are mapped into a single octave. This gives a measure of pitch accuracy ignoring the octave errors.

4. Results

Our proposed DA-Net is a model with great improvement based on the CREPE structure. Here, we compared our proposed DA-Net with CREPE and explored two kinds of single attention networks and the DA-Nets with three kinds of DA-Modules. According to the experimental results, the DA-Net with a tightly coupled mode achieved the best performances. We also compared our proposed DA-Net with three previous methods.

4.1. Comparison of Single or Dual Attention Networks

The CREPE model reported in [26] was trained on the MIR-1K, Bach10, RWC-synth, MedleyDB, MDB-stem-synth, and Nsynth datasets. These six datasets are all composed of vocal and instrumental audio. The CREPE method is, therefore, expected to work best on this type of signal. We retrained the CREPE models on the iKala [46] and MDB-stem-

synth [47] datasets. Table 2 shows the results of our proposed networks with different attention mechanisms on the iKala dataset.

Table 3 shows the results of the same networks as that in Table 2 on the MDB-stem-synth dataset. CREPE in Tables 2 and 3 denotes that we trained the CREPE model with a six-layer CNN only on the iKala dataset and only on the MDB-stem-synth dataset, respectively. The EA denotes the network using the element-wise attention module instead of the DA-Module in the proposed network. The CA denotes that in the architecture of DA-Net, the channel-wise attention module was instead of the DA-Module. The DA-S denotes our proposed DA-Net with Serial mode. In each DA-S module, the inputs are fed into the EA block, and its outputs are further fed into the CA block.

DA-P denotes that each DA-Module was with Parallel mode in the proposed DA-Net. Each DA-P module includes two branches, where the top half branch applies CNN to extract the feature maps of inputs before sending it to the CA block. The lower half branch is the EA block. DA-TC denotes that the DA-Net was with the Tightly Coupled mode. In the top half part of the EA block of DA-TC modules, the feature maps after a 1-dimension CNN are fed into the CA block to learn the attention of various channels, meanwhile, it served as the feature maps of the EA block.

From the results in Tables 2 and 3, we can see that the scores of the network were quite different on the two datasets. On the two datasets, the average scores of the EA, CA, DA-S, DA-P, and DA-TC were superior to the CREPE model. On the iKala dataset, DA-TC outperformed CREPE by 0.72% at RPA and 0.78% at RCA. On the MDB-stem-synth dataset, the scores of EA were higher than that of CREPE by 4.85% (RPA) and 3.31% (RCA). The scores of CA and DA-S were also slightly higher than those of the CREPE. It can be seen in Section 2.2 that DA-TC had six fewer convolution operations than DA-P.

Compared with DA-P, DA-TC had not only fewer parameters but also higher performance. DA-TC showed 5.60% and 3.88% improvements at RPA and RCA on MDB-stem-synth, respectively. Therefore, the DA-TC proposed in this paper maximized the effect of the element-wise attention mechanism and channel-wise attention mechanism to improve the estimation performance. Figure 6 illustrates the estimated contour by DA-TC drawing on the spectrum of a clip in the MDB-stem-synth dataset and the comparison with the ground truth.

Table 2. Comparison of our proposed network with different attention mechanisms on the iKala dataset.

Structure	RPA	RCA
CREPE [26]	91.36%	91.48%
EA	91.46%	91.62%
CA	91.71%	91.92%
DA-S	91.91%	92.07%
DA-P	91.68%	91.83%
DA-TC	**92.08%**	**92.26%**

Table 3. Comparison of our proposed network with different attention mechanisms on the MDB-stem-synth dataset.

Structure	RPA	RCA
CREPE [26]	87.55%	89.53%
EA	92.40%	92.84%
CA	88.77%	89.63%
DA-S	90.30%	91.96%
DA-P	91.31%	91.76%
DA-TC	**93.15%**	**93.41%**

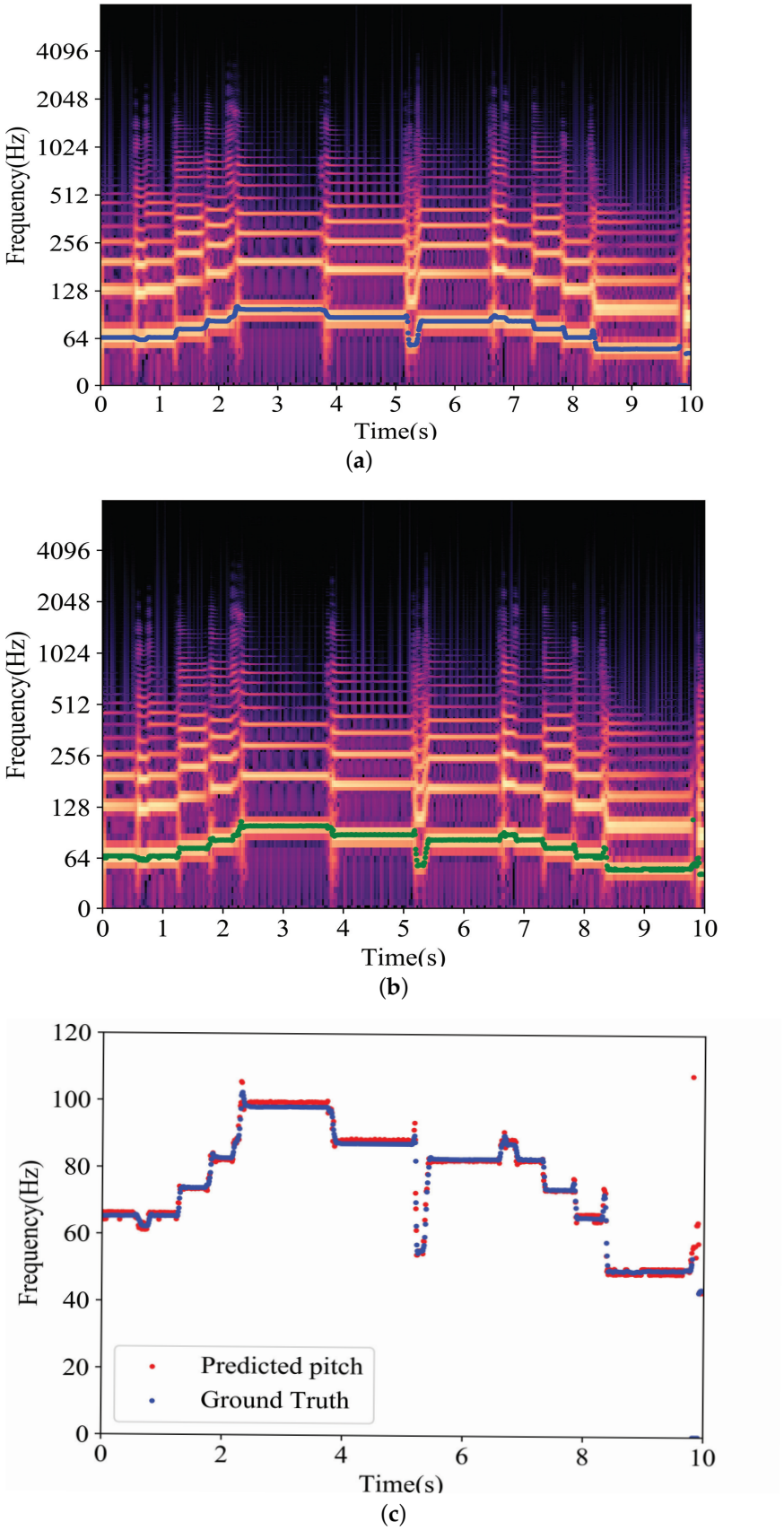

Figure 6. Pitch contour of a clip in the MDB-stem-synth dataset. (**a**) The ground truth pitch contour drawing on the spectrum with log-scale frequency. (**b**) Predicted pitch contour drawing in the same spectrum as in (**a**). (**c**) Curves of the predicted pitch and ground truth.

4.2. Comparison with Previous Methods

The proposed DA-TC model for pitch estimation was also compared with the pYIN [9], SWIPE [11] and SPICE [29] algorithms. pYIN and SWIPE are the classic traditional pitch estimators, and SPICE is based on a neural network trained in a self-supervised way. Table 4 shows the results of pitch estimation from different methods on the MDB-stem-synth dataset.

The results of pYIN and SWIPE were listed according to the reported results in the paper [26]. In the reference [29], the authors only reported the results of RPA but without that of RCA. The results in Table 4 show that the average score of DA-TC was higher than pYIN, SWIPE, and SPICE by 1.15%, 0.65%, and 4.05% in the term of RPA.

Table 4. Comparison of the proposed method (DA-TC) and three methods on the MDB-stem-synth dataset.

Structure	RPA	RCA
pYIN [9]	91.90%	93.60%
SWIPE [11]	92.50%	93.60%
SPICE [29]	89.10%	–
DA-TC	**93.15%**	93.41%

5. Conclusions

In this paper, we proposed a data-driven method, DA-Net, for pitch estimation of monophonic music operating on raw samples. We investigated our proposed DA-Net with three kinds of combination modes and the single attention networks, which included EA and CA. The iKala and MDB-stem-synth datasets were used to train and test. According to the experimental results, we can draw our conclusions as follows:

(i) The performances of the models when introducing single attention mechanisms, such as EA and CA, were better than CREPE.
(ii) Dual attention networks (DA-S, DA-P, and DA-TC) outperformed the CREPE model. Among them, DA-TC, which maximizes the effect of the element-wise attention mechanism and the channel-wise attention mechanism, achieved the best performance.
(iii) DA-TC obtained the best results and outperformed pYIN, SWIPE, and SPICE on the MDB-stem-synth dataset.

Author Contributions: Conceptualization, Y.H. and W.M.; methodology , Y.H. and W.M.; software, W.M.; validation, W.M. and Y.H.; resources, Y.H. and H.H.; writing—original draft preparation, W.M. and Y.H.; supervision, Y.H. and H.H.; funding acquisition, Y.H. and H.H. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Natural Science Foundation of China (NSFC) (61761041, U1903213). The Funds for Creative Research Groups of Higher Education of Xinjiang Uygur Autonomous Region under Grant No.XJEDU2017T002.

Data Availability Statement: http://mac.citi.sinica.edu.tw/ikala/ (accessed on 15 April 2021); https://sites.google.com/site/unvoicedsoundseparation/mir-1k (accessed on 15 April 2021).

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

References

1. Bittner, R.M.; Salamon, J.; Tierney, M.; Mauch, M.; Cannam, C.; Bello, J.P. Medleydb: A multitrack dataset for annotation-intensive mir research. *ISMIR* 2014,*14*, 155–160.
2. Bosch, J.; Gómez, E. Melody extraction in symphonic classical music: A comparative study of mutual agreement between humans and algorithms. In Proceedings of the 9th Conference on Interdisciplinary Musicology—CIM14, Berlin, Germany, 4–6 December 2014.

3. Mauch, M; Cannam, C; Bittner, R; Fazekas, G; Salamon, J; Dai, J; Bello, J; Dixon, S. Computer-aided melody note transcription using the Tony software: Accuracy and efficiency. **2015**. Available online: https://qmro.qmul.ac.uk/xmlui/handle/123456789/7247 (accessed on 15 April 2021).
4. Rodet, X. Synthesis and processing of the singing voice. In Proceedings of the 1st IEEE Benelux Workshop on Model based Processing and Coding of Audio (MPCA-2002), Leuven, Belgium, 15 November 2002; pp. 15–31.
5. Downie, J.S. Music information retrieval. *Annu. Rev. Inf. Sci. Technol.* **2003**, *37*, 295–340.
6. Panteli, M.; Bittner, R.; Bello, J.P.; Dixon, S. Towards the characterization of singing styles in world music. In Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 5–9 March 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 636–640.
7. Klapuri, A. Automatic transcription of music. Master's Thesis, Tampere University of Technology, Tampere, Finland, April 1998.
8. De Cheveigné, A.; Kawahara, H. YIN, a fundamental frequency estimator for speech and music. *J. Acoust. Soc. Am.* **2002**, *111*, 1917–1930.
9. Mauch, M.; Dixon, S. pYIN: A fundamental frequency estimator using probabilistic threshold distributions. In Proceedings of the 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Florence, Italy, 4–9 May 2014; IEEE: Piscataway, NJ, USA, 2014; pp. 659–663.
10. Camacho, A.; Harris, J.G. A sawtooth waveform inspired pitch estimator for speech and music. *J. Acoust. Soc. Am.* **2008**, *124*, 1638–1652.
11. Zahorian, S.A.; Hu, H. A spectral/temporal method for robust fundamental frequency tracking. *J. Acoust. Soc. Am.* **2008**, *123*, 4559–4571.
12. Hua, K. Nebula: F0 estimation and voicing detection by modeling the statistical properties of feature extractors. *arXiv* **2017**, arXiv:1710.11317.
13. Zhang, J.; Tang, J.; Dai, L.-R. RNN-BLSTM Based Multi-Pitch Estimation. In Proceedings of the Interspeech, San Francisco, CA, USA, 8–12 September 2016; pp. 1785–1789.
14. Gonzalez, S.; Brookes, M. PEFAC-a pitch estimation algorithm robust to high levels of noise. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2014**, *22*, 518–530.
15. Verma, P.; Schafer, R.W. Frequency Estimation from Waveforms Using Multi-Layered Neural Networks. In Proceedings of the Interspeech, San Francisco, CA, USA, 8–12 September 2016; pp. 2165–2169.
16. Han, K.; Wang, D.L. Neural network based pitch tracking in very noisy speech. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2014**, *22*, 2158–2168.
17. Bittner, R.M.; McFee, B.; Salamon, J.; Li, P.; Bello, J.P. Deep Salience Representations for F0 Estimation in Polyphonic Music. In Proceedings of the ISMIR, Suzhou, China, 23–27 October 2017; pp. 63–70.
18. Bittner, R.M.; McFee, B.; Bello, J.P. Multitask learning for fundamental frequency estimation in music. *arXiv* **2018**, arXiv:1809.00381.
19. Basaran, D.; Essid, S.; Peeters, G. Main melody extraction with source-filter nmf and crnn. In Proceedings of the 19th International Society for Music Information Retreival, Paris, France, 23–27 September 2018.
20. Doras, G.; Esling, P.; Peeters, G. On the use of u-net for dominant melody estimation in polyphonic music. In Proceedings of the 2019 International Workshop on Multilayer Music Representation and Processing (MMRP), Milano, Italy, 24–25 January 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 66–70.
21. Lu, W.T.; Su, L. Vocal Melody Extraction with Semantic Segmentation and Audio-symbolic Domain Transfer Learning. In Proceedings of the ISMIR, Paris, France, 23–27 September 2018; pp. 521–528.
22. Chen, M.-T.; Li, B.-J.; Chi, T.-S. Cnn based two-stage multi-resolution end-to-end model for singing melody extraction. In Proceedings of the ICASSP 2019—2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 1005–1009.
23. Hsieh, T.-H.; Su, L.; Yang, Y.-H. A streamlined encoder/decoder architecture for melody extraction. In Proceedings of the ICASSP 2019—2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 156–160.
24. Xu, S.; Shimodaira, H. Direct F0 Estimation with Neural-Network-Based Regression. In Proceedings of the Interspeech, Graz, Austria, 15–19 September 2019; pp. 1995–1999.
25. Airaksinen, M.; Juvela, L.; Alku, P.; Räsänen, O. Data Augmentation Strategies for Neural Network F0 Estimation. In Proceedings of the ICASSP 2019—2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 6485–6489.
26. Kim, J.W.; Salamon, J.; Li, P.; Bello, J.P. Crepe: A convolutional representation for pitch estimation. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 161–165.
27. Ardaillon, L.; Roebel, A. Fully-convolutional network for pitch estimation of speech signals. In Proceedings of the Insterspeech 2019, Graz, Austria, 15–19 September 2019.
28. Dong, M.; Wu, J.; Luan, J. Vocal Pitch Extraction in Polyphonic Music Using Convolutional Residual Network. In Proceedings of the Insterspeech 2019, Graz, Austria, 15–19 September 2019; pp. 2010–2014.

29. Gfeller, Beat and Frank, Christian and Roblek, Dominik and Sharifi, Matt and Tagliasacchi, Marco and Velimirović, Mihajlo. Pitch Estimation Via Self-Supervision. In Proceedings of the ICASSP 2020—2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 3527–3531.
30. Dauphin, Y.N.; Fan, A.; Auli, M.; Grangier, D. Language modeling with gated convolutional networks. In Proceedings of the International Conference on Machine Learning, Sydney, Australia, 6–11 August 2017; PMLR: Mountain View, California, USA, 2017; pp. 933–941.
31. Tan, K.; Wang, D.L. Learning complex spectral mapping with gated convolutional recurrent networks for monaural speech enhancement. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2019**, *28*, 380–390.
32. Tan, K.; Chen, J.; Wang, D.L. Gated residual networks with dilated convolutions for monaural speech enhancement. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2018**, *27*, 189–198.
33. Shi, Z.; Lin, H.; Liu, L.; Liu, R.; Han, J.; Shi, A. Deep Attention Gated Dilated Temporal Convolutional Networks with Intra-Parallel Convolutional Modules for End-to-End Monaural Speech Separation. In Proceedings of the Interspeech, Graz, Austria, 15–19 September 2019; pp. 3183–3187.
34. Geng, H.; Hu, Y.; Huang, H. Monaural Singing Voice and Accompaniment Separation Based on Gated Nested U-Net Architecture. *Symmetry* **2020**, *12*, 1051.
35. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE conference on computer vision and pattern recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.
36. Wu, Y. Deep convolutional neural network based on densely connected squeeze-and-excitation blocks. *AIP Adv.* **2019**, *9*, 065016.
37. Gu, J.; Sun, X.; Zhang, Y.; Fu, K.; Wang, L. Deep residual squeeze and excitation network for remote sensing image super-resolution. *Remote. Sens.* **2019**, *11*, 1817.
38. Park, Y.J.; Tuxworth, G.; Zhou, J. Insect classification using Squeeze-and-Excitation and attention modules-a benchmark study. In Proceedings of the 2019 IEEE International Conference on Image Processing (ICIP), Taipei, Taiwan, 22–25 September 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 3437–3441.
39. Wu, J.; Li, Q.; Liang, S.; Kuang, S.-F. Convolutional Neural Network with Squeeze and Excitation Modules for Image Blind Deblurring. In Proceedings of the 2020 Information Communication Technologies Conference (ICTC), Nanjing, China, 29–31 May 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 338–345.
40. Fu, J.; Liu, J.; Tian, H.; Li, Y.; Bao, Y.; Fang, Z.; Lu, H. Dual attention network for scene segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 3146–3154.
41. Yu, Shuai and Sun, Xiaoheng and Yu, Yi and Li, Wei. Frequency-temporal attention network for singing melody extraction. In Proceedings of the ICASSP 2021—-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 6–12 June 2021; pp. 251–255.
42. Hu, Z.; Luo, Y.; Lin, J.; Yan, Y.; Chen, J. Multi-Level Visual-Semantic Alignments with Relation-Wise Dual Attention Network for Image and Text Matching. In Proceedings of the IJCAI 2019, Macao, 10–16 August 2019; pp. 789–795.
43. Li, B.; Ye, W.; Sheng, Z.; Xie, R.; Xi, X.; Zhang, S. Graph enhanced dual attention network for document-level relation extraction. In Proceedings of the 28th International Conference on Computational Linguistics, Barcelona, Spain, 13–18 September 2020; pp. 1551–1560.
44. Zhu, Y.; Zheng, W.; Tang, H. Interactive dual attention network for text sentiment classification. *Comput. Intell. Neurosci.* **2020**, *2020*, 8858717.
45. Wan, J.; Xie, Z.; Xu, Y.; Chen, S.; Qiu, Q. DA-RoadNet: A Dual-Attention Network for Road Extraction from High Resolution Satellite Imagery. *IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens.* **2021**, *14*, 6302–6315.
46. Chan, T.-S.; Yeh, T.-C.; Fan, Z.-C.; Chen, H.-W.; Su, L.; Yang, Y.-H.; Jang, R. Vocal activity informed singing voice separation with the iKala dataset. In Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Queensland, Australia, 19–24 April 2015; IEEE: Piscataway, NJ, USA, 2015; pp. 718–722.
47. Salamon, J.; Bittner, R.M.; Bonada, J.; Bosch, J.J.; Gómez Gutiérrez, E.; Bello, J.P. An analysis/synthesis framework for automatic f0 annotation of multitrack datasets. In Proceedings of the ISMIR 2017 Proceedings of the 18th International Society for Music Information Retrieval Conference, Suzhou, China, 23–27 October 2017.
48. Kingman, D.P..; Ba, J. Adam: A Method for Stochastic Optimization. Conference paper. In Proceedings of the 3rd International Conference for Learning Representations, San Diego, CA, USA, 7–9 May 2015.
49. Ioffe, S.; Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In Proceedings of the International Conference on Machine Learning, Lille, France, 6–11 July 2015; pp. 448–456.
50. Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R. Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **2014**, *15*, 1929–1958.
51. Raffel, C.; McFee, B.; Humphrey, E.J.; Salamon, J.; Nieto, O.; Liang, D.; Ellis, D.P.W.; Raffel, C.C. mir_eval: A transparent implementation of common MIR metrics. In Proceedings of the 15th International Society for Music Information Retrieval Conference, ISMIR 2014, Taipei, Taiwan, 27–31 October 2014.