Every Little Bit Helps: Exploring Better Utilization of Unlabeled Data for Semi-supervised Singing Melody Extraction Using Multi-bands Diffusion Model

Shuai Yu[1], **Xiaoliang He**[2], **Kangjie Dong**[2], **Yi Yu**[*3]

[1] School of Information and Communication Engineering, Dalian University of Technology
[2]School of Information and Intelligent Science, Donghua University
[3]Graduate School of Advanced Science and Engineering, Hiroshima University

Abstract

Semi-supervised singing melody extraction (SSME) is one of the key tasks in the field of music information retrieval (MIR). Recently, several SSME methods have been proposed and achieved remarkable successes. However, existing methods are still facing two critical issues: firstly, there is a lack of an effective data augmentation method for SSME, which results in insufficient utilization of unlabeled data. Secondly, existing SSME methods discards too much unlabeled data in the stage of consistency regularization, which hinders the further improvements of SSME task. In this paper, we present *ELH-SME*, a novel framework that better utilizes the unlabeled musical data for SSME task. Specifically, our proposed ELH-SME framework consists of three modules: (1) we first propose a diffusion-based multi-bands augmentation (DMA) method to increase the amounts of training data. The proposed DMA methods employs a diffusion-based model to generate perturbation at the specific frequency bands in an end-to-end manner, thereby avoiding sharply perturbations to the spectrogram. (2) To improve the utilization rate of unlabeled data, we suggest a global-class confidence (GCC) module. During the phase of consistency regularization, we consider both the global-wise and class-wise confidence values, improving the utilization rate of unlabeled data. (3) To further improve the utilization of unlabeled data, we also propose to enhance the representation capability of unlabeled data by extracting channel-level features from labeled data via channel cross-attention (CCA). We evaluate our proposed framework on several well-known public available datasets, and the conducted experiments demonstrate the effectiveness of our method.

Introduction

Singing melody extraction (SME) is one of the challenging tasks in the field of music information retrieval (MIR) (Salamon et al. 2014). It aims to extract fundamental frequency (`f0`) from the polyphonic music. The extracted `f0` can be employed as acoustic features for many downstream MIR applications, such as query-by-humming (Wang and Jang 2015), cover song identification (Serra, Gómez, and Herrera 2010) and music recommendation (Knees and Schedl 2015). However, compared with other tasks in MIR, SME needs more expensive pixel-level annotations for training. Thus, semi-supervised singing melody extraction (SSME) has become an important task that utilizes both labeled and unlabeled musical data (Kum et al. 2020).

With the development of deep learning (DL) technique, many DL based SME methods have been proposed and achieved remarkable successes (Liu et al. 2025; Wang et al. 2025; Hu et al. 2025). Recently, to alleviate the scarcity of labeled data, SSME has become an active research direction. A number of SSME methods have been proposed and achieved significant results (Kum et al. 2020; Yu 2024; He et al. 2025). Despite the successes, SSME task is still facing two critical issues: firstly, there is a lack of an effective data augmentation method for SSME, which results in insufficient utilization of unlabeled data. It has been claimed by prior works (Yu 2024) that SSME task is very sensitive to the data augmentation, in this paper, called ***sensitivity issue***. Since data augmentation is a critical part in the semi-supervised learning (Berthelot et al. 2019; Olsson et al. 2021), the sensitivity issue limits the SSME task to apply effective data augmentation methods for better utilization of the unlabeled data. Secondly, existing SSME methods discard too much unlabeled data in the stage of consistency regularization, which hinders the further improvements of SSME task. As shown in Fig.1, existing SSME methods discard about 50% (or more) unlabeled data in the stage of consistency regularization.

To alleviate the sensitivity issue, A pioneer work, MC-SSME (Yu 2024), proposed to apply data augmentation on the raw waveform data (i.e., time-domain) and then extracted singing melody in the spectrogram (i.e., frequency-domain). However, it is challenging to guarantee that data augmentations applied in the time-domain preserve their efficacy in the subsequent frequency-domain melody extraction. HKDSME (Yu et al. 2024a) proposed to use harmonic information in the spectrogram as an additional supervision signal to better utilize the unlabeled data. Unfortunately, HKDSME does not include a data augmentation process, though it alleviates the sensitivity issue. A natural question is raised: *How to design an effective data augmentation method that works well for SSME?*

Recently, Denoising Diffusion Probability Model (DDPM) has shown impressive abilities on generation tasks, such as image generation (Rombach et al. 2022;

*Yi Yu is the corresponding author.

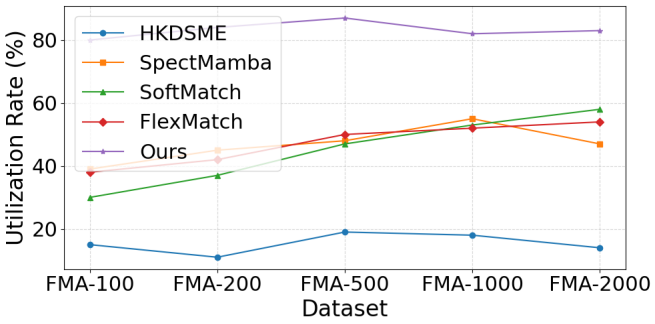


Figure 1: The utilization rates of unlabeled data using different models. The X-axis shows the different amounts of unlabeled data. The Y-axis shows the utilization rate of unlabeled data using different models.

Brooks, Holynski, and Efros 2023) and music generation (Li et al. 2024; Wang, Min, and Xia 2024). Inspired by the powerful generation ability of DDPM, we are motivated to regard the data augmentation in the frequency-domain as a generative task so that the perturbations in the spectrogram could be generated in an end-to-end manner, thereby avoiding sharply perturbations to the spectrogram. In other words, we hope the generative model (e.g., DDPM) could learn how to perform good data augmentation in the frequency-domain for SSME task. The straight-forward solution is to directly apply DDPM for generating augmentation. However, it is not practical due to two reasons: firstly, in the spectrogram, the amplitude value distributions varies greatly. A fine-grained data augmentation method is needed to adapt the various distributions. For example, in popular music, the amplitude values are concentrated in the low and medium frequency bands. And in opera music, the amplitude values are concentrated in the medium and high frequency bands. Thus, it is not practical to directly use DDPM to perform data augmentations on various musical data. Secondly, the singing voice will generate larger amplitude values in the higher frequency bands than in `f0` (Salamon et al. 2014), which will mislead the final prediction. Thus, information should be shared between different frequency bands to avoid the augmented spectrograms from misleading the final predictions.

On the other hand, consistency regularization is widely used in the SSME task (Yu et al. 2024a; He et al. 2025). However, the utilization of unlabeled data is relatively low in the stage consistency regularization. In the field of semi-supervised learning, prior research works focused on using dynamic confidence thresholds to select unlabeled data for accurate predictions. FlexMatch (Zhang et al. 2021a) proposed a method that computes the learning difficulty of each class, improving the utilization of the unlabeled data. However, since this method does not include a global confidence for unlabeled data, some samples with small confidence values (e.g., close to zero) will still be trained rather than discard. SoftMatch (Chen et al. 2023) proposed to use a function to assign weight to unlabeled data. The weighted value stands for the difference between sample and global confidence. However, this method does not establish the connection between global-wise and class-wise confidences.

Following the above analysis, in this work, we propose a diffusion-based multi-bands augmentation (DMA) method to increase the amounts of training data in an end-to-end manner. Specifically, the proposed DMA method employs a diffusion-based model to generate perturbations at different frequency bands (i.e., low, medium and high frequency bands) separately, thereby adapting various amplitude distributions through such a fine-grained fashion. To alleviate the interference of large amplitude values in the higher frequency bands, we then fuse the separated augmented outputs to share the separated information at different frequency bands and further obtain the final augmented spectrogram.

Then, we propose a global-class confidence (GCC) module. During the phase of consistency regularization, the proposed GCC module enables the model to adaptively control the global-wise and class-wise thresholds for unlabeled samples, improving the utilization rate of unlabeled data. To further improve the utilization rate of unlabeled data, we also propose a channel cross-attention (CCA) module that extracts channel-wise features from unlabeled musical audio, enhancing the representation capability of labeled data.

The contribution of this paper is summarized as follows:

- To alleviate the sensitivity issue, we propose a novel diffusion-based multi-bands augmentation (DMA) module to increase the amounts of training data. The proposed DMA module first generate perturbations at different frequency bands separately and then the separated perturbations are fused to obtain the final augmented spectrogram.
- To improve the utilization rate of the unlabled data in SSME task, we then propose a global-class confidence (GCC) module. During the stage of consistency regularization, the proposed GCC module enables the model to adaptively control the threshold for unlabeled samples, improving the utilization rate of unlabeled data.
- To further improve the utilization rate of unlabeled data, we also propose a channel cross-attention (CCA) module that extracts channel-wise features from unlabeled musical audio, enhancing the representation capability of labeled data.
- We use MIR-1K dataset and part of music tracks of the Medley DB dataset as labeled data for training the model and we evaluate the performance on the well-known ADC2004, MIREX 05, iKala and another part of Medley DB. The experimental results demonstrate the superiority of our method compared with other state-of-the-art ones.

Related Work

Singing Melody Extraction

Many data-driven based approaches have been proposed for SME (Kum, Oh, and Nam 2016; Lu, Su et al. 2018; Su 2018; Chen, Li, and Chi 2019). In addition, the use of musical prior knowledge and structural priors has further inspired the design of melody extraction models (Park and Yoo 2017; Chou, Chen, and Chi 2018; Kum and Nam 2019). The

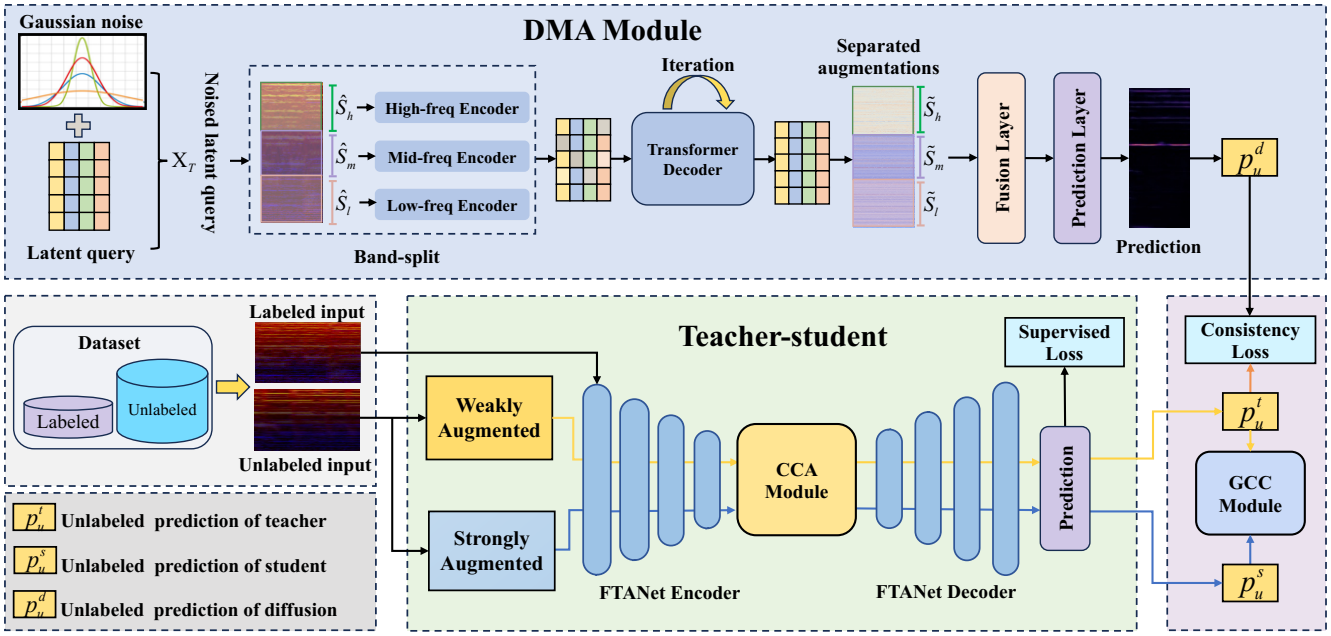


Figure 2: The framework of our proposed *ELH-SME*. The proposed ELH-SME framework consists of three modules: (a) diffusion-based multi-bands augmentation (DMA) module, (b) global-class confidence (GCC) module and (c) channel cross-attention (CCA) module.

relationship between frequencies can be further captured through multi-dilation or attention networks (Gao, You, and Chi 2020; Du et al. 2021; Yu et al. 2023, 2024b; Yu, He, and Zhang 2024), or harmonic constant-Q transform (HCQT) (Bittner et al. 2017). Recently, there are are a number of research works (Yu et al. 2021b; Yu 2024; Yu et al. 2024a, 2025) have been proposed for SSME. Unfortunately, most of the above methods overlook the utilization of unlabeled data. In this work, we focus on alleviating the sensitivity issue and improve the utilization rate of unlabeled data in SSME task.

Diffusion Models

Diffusion model is related to our work, which is initially proposed by (Sohl-Dickstein et al. 2015). In the field of audio information processing, there are many research works propose to apply diffusion models to sound event detection (Bhosale et al. 2024), music generation (Li et al. 2024; Wang, Min, and Xia 2024) and other audio-based applications (Xu et al. 2024; Evans et al. 2024). Very recently, Azizi et al. (Azizi et al. 2023) proposed to use text-to-image diffusion models to synthesis training data for image classification. Islam et al. (Islam et al. 2024) proposed a image-mixing data augmentation method for image classification. Different from prior works, in this work, we propose to use diffusion models to generate perturbations at specific frequency bands in the single spectrogram, and then fuse the separated augmentations to generate the final spectrogram.

Consistency Regularization

In the field of machine learning, consistency regularization methods have become a mainstream research direction for semi-supervised learning (Jeong and Shin 2020; Zhang et al. 2021b; Saito, Kim, and Saenko 2021; Cheng et al. 2022; Ni and Koniusz 2023). A number of consistency regularization methods (Laine and Aila 2017; Tarvainen and Valpola 2017; Athiwaratkun et al. 2019; Berthelot et al. 2019; Xie et al. 2020) have been proposed to select reliable unlabeled data. In the task of SSME, researchers have proposed several efficient methods to improve the accuracy of pseudo labels. However, the existing consistency regularization methods in SSME overlook the utilization rate of unlabeled data, thereby achieving suboptimal results. In this paper, we focus on improving the utilization rate of unlabeled data, thereby improving the performances of SSME task.

Method

The overview of the proposed framework is presented in Fig.2. We choose to employ FTANet (Yu et al. 2021a) as the backbone of ELH-SME. A diffusion-based multi-bands augmentation (DMA) module is proposed to generate perturbations at the specific frequency bands separately and then fuse the separated augmentations to share the information at different frequency bands. Subsequently, the proposed global-class confidence (GCC) module learns global-wise and class-wise confidence values to improve the utilization rate of unlabeled data. Finally, we devise a channel cross-attention (CCA) module to extract rich features from unlabeled data for improving the representation learning of unlabeled data. We first introduce the semi-supervised learn-

ing setting. Next, we will introduce each component in the following subsections.

Semi-supervised Learning Setup

Our work is under the SSME setting, the inputs are from both labeled and unlabeled data. For the input data, the music signal can be denoted as $D = \{D_l, D_u\}$. $D_l = \{(x_1, y_1), (x_2, y_2), ..., (x_m, y_M)\}$ and $D_u = \{u_1, u_2, ..., u_N\}$ denote the labeled music data and unlabeled music data, respectively. M and N are the number of labeled and unlabeled data. T denotes the whole training dataset. We employ a teacher-student architecture to build the semi-supervised learning. Following MCSSME (Yu 2024), we apply time-domain data augmentation to the unlabeled data as the input of teacher-student model. The learning objective function is constructed in the following form:

$$\min_{\theta}\{\mathcal{L}_l(D_l, \theta) + \omega\mathcal{L}_{CL}(D_u, \theta)\}, \tag{1}$$

where $\mathcal{L}_l$ is the loss function for supervised learning and $\mathcal{L}_{CL}$ is the consistency loss for unsupervised learning. ω is a non-negative parameter, and θ represents the parameters of our proposed framework.

Diffusion-based Multi-bands Augmentation for SSME

Distinct from existing SSME models (Kum et al. 2020; Yu 2024), we aim to augment the unlabeled music data in the frequency domain. Our goal is to augment the unlabeled spectrogram through a generative model (e.g., DDPM) in an end-to-end manner. To achieve this, we first employ a diffusion-based model to generate fine-grained perturbations at different frequency bands to adapt the various amplitude distribution. Then, we fuse the separated augmentations to obtain the final augmented spectrogram.

Given an input spectrogram $\mathbf{S} \in \mathbb{R}^{F\times T}$, where F, T denote frequency bands and time steps, respectively[1]. We first generate the noised latent query for thereafter diffusion process with $\mathbf{S}$. Specifically, we perform element-wise addition between Gaussian noise and a randomly initialized learnable matrix $\mathbf{Q} \in \mathbb{R}^{F\times T}$, called latent query. Then the noised latent query is added to the input spectrogram $\mathbf{S}$ to obtain a noised input spectrogram $\hat{\mathbf{S}}$. Next, we split $\hat{\mathbf{S}}$ into three parts along the frequency bands: $\hat{\mathbf{S}}_h$, $\hat{\mathbf{S}}_m$, and $\hat{\mathbf{S}}_l$, where $\hat{\mathbf{S}}_h$, $\hat{\mathbf{S}}_m$, $\hat{\mathbf{S}}_l$ denote the divided spectrograms at the high-, medium-, and low- frequency bands from $\hat{\mathbf{S}}$, respectively. When obtaining the divided spectrograms, we feed them into the DDPM model to generate the separated augmentations to the three spectrograms:

$$\begin{aligned}\tilde{\mathbf{S}}_h &= DDPM(\hat{\mathbf{S}}_h),\\ \tilde{\mathbf{S}}_m &= DDPM(\hat{\mathbf{S}}_m),\\ \tilde{\mathbf{S}}_l &= DDPM(\hat{\mathbf{S}}_l).\end{aligned} \tag{2}$$

[1]Here, we omit the channel dimension for simplification. We use CFP representation as the input spectrogram, and there are 3 channels actually.

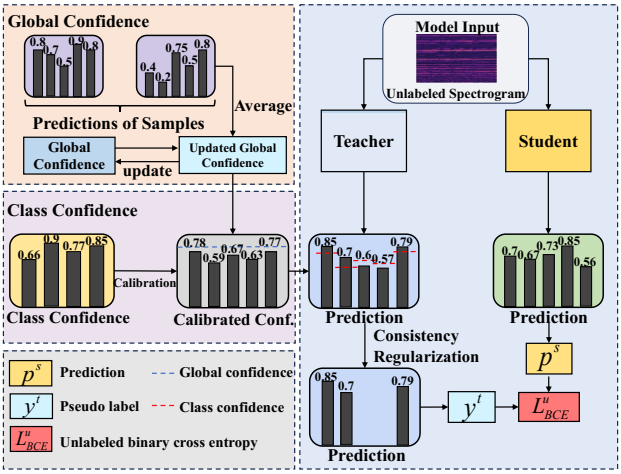


Figure 3: Detailed architecture of the proposed global-class confidence module. The left part illustrates the calculation of global-wise and class-wise confidence values. The right part shows how the calculated confidence is used in SSME.

Since the singing voice generates higher amplitude values at the higher frequency bands, we feed the separated augmentations $\tilde{\mathbf{S}}_h, \tilde{\mathbf{S}}_m, \tilde{\mathbf{S}}_l$ into a fusion layer to obtain the final augmentation $\mathbf{S}_{aug}$:

$$\mathbf{S}_{aug} = Fusion(\tilde{\mathbf{S}}_h, \tilde{\mathbf{S}}_m, \tilde{\mathbf{S}}_l), \tag{3}$$

and the fusion layer is composed of a three-layer multilayer perceptron (MLP) to fuse the information from different frequency bands. Finally, we calculate the consistency loss $\mathcal{L}_{CL}$ for unlabeled data:

$$\mathcal{L}_{CL} = KL(p_u^d, p_u^t) + KL(p_u^s, p_u^t), \tag{4}$$

where KL is the Kullback-Leibler divergence, p_u^d is the prediction using the augmented spectrogram $\mathbf{S}_{aug}$, p_u^s is the prediction of student, and p_u^t is the prediction of teacher.

Global-Class Confidence for SSME

To improve the utilization rate of unlabeled data in SSME, we propose the global-class confidence (GCC) module to better utilize the unlabeled music data as shown in Fig.3. The key idea of GCC module is to take consideration of both global-wise and class-wise confidences so that we can use more unlabeled data for training, thereby improving the utilization rate of unlabeled data. To this end, we first computes the average confidence in each epoch, and we take the average confidence as the global-wise confidence C_g :

$$C_g = \frac{1}{N}\sum_{i=1}^{N} c_i \tag{5}$$

where c_i is the confidence value of unlabeled sample u_i. The C_g is updated during training. After obtaining the global-wise confidence, we then similarly compute the class-wise confidence C_l.

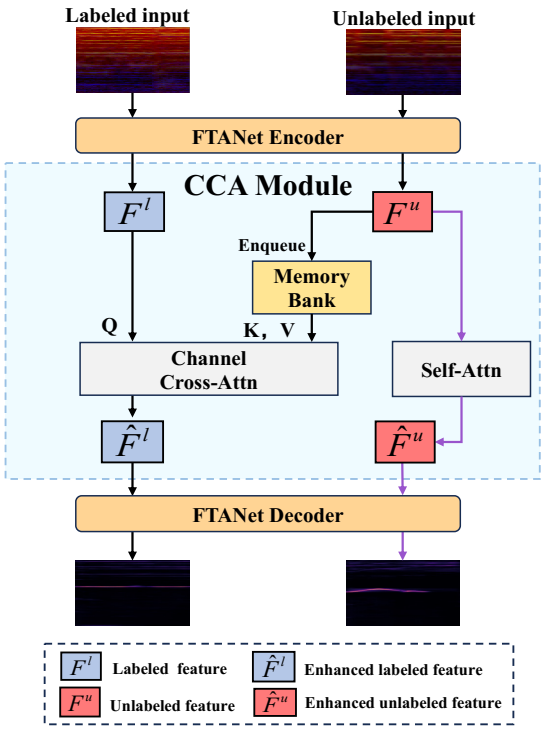


Figure 4: Detailed architecture of the proposed channel cross-attention module.

In order to simultaneously exploit the global-wise and class-wise confidences, a straight-forward solution is to linearly combine the global-wise and class-wise confidences. However, such combination obtains suboptimal performances. We find that there is a large difference between global-wise and class-wise confidence values. Moreover, the global-wise and class-wise confidence values are dynamic in every epoch. It is difficult to set a hyperparameter that balances the two parts.

Consequently, we devise a novel calibration-based method, which is more practical for SSME. The core idea is to take the class-wise confidence value as the weight of the global-wise confidence. To be specific, we first compute the class-wise confidence C_l as mentioned above, then we normalize C_l to a value $\in (0, 1)$:

$$C_l^k = \frac{C_l^k}{\sum_k C_l^k} \tag{6}$$

where C_l^k denotes the confidence value of k-th class. Next, we compute the confidence value: $C^k = C_l^k * C_g$. Finally, we select the unlabeled data whose confidence value is larger than C^k for training.

Channel Cross-Attention for SSME

To further improve the utilization of unlabeled data, we also propose a channel cross-attention (CCA) module to extract channel-wise features from unlabeled data to enhance the representation capability of labeled data. The goal of CCA module is to alleviate the scarcity of labeled data in SSME, and fully utilize the information in the unlabeled musical data. Specifically, we first employ a memory bank $B \in \mathbb{R}^{K \times d}$ to store the representative features for each class [2], where K denotes the number of classes and d is the feature dimension. We then employ the cosine similarity to calculate the similarity between the extracted feature $F^u \in \mathbb{R}^d$ (from unlabeled sample u_i) and the representative feature B_i in B:

$$sim(F^u, B_i) = cos(F^u, B_i) = \frac{(F^u)^T B_i}{|F^u||B_i|} \tag{7}$$

Next, we use the calculated similarities to predict the class of u_i. We then perform channel cross-attention to obtain features $\mathcal{F}$ from unlabeled samples:

$$\mathcal{F} = CCA(u_i, B_i) = \sum_j \alpha_{ij} u_{ij} \tag{8}$$

$$\alpha_{ij} = \frac{\phi(B_i, u_{ij})}{\sum_k \phi(B_i, u_{ik})} \tag{9}$$

where u_{ij} is the j-th channel-wise feature of u_i and $\phi(\cdot, \cdot)$ is a three-layer MLP to calculate the degree of matching:

$$\phi(B_i, u_{ij}) = MLP(W_1 B_i + W_2 u_{ij}) \tag{10}$$

where W_1 and W_2 are the learnable parameters. In addition, we also employ self-attention technique (Vaswani et al. 2017) to enhance the channel-wise feature for the unlabeled data. Finally, the enhanced features are fed into the decoder of FTANet to obtain the final prediction of singing melody as shown in Fig.4.

	Dataset	# of Tracks	Duration
Training (Labeled)	MIR-1K	1000	2h 13min
	Medley DB	30	1h 58min
Training (Unlabeled)	FMA	2000	15h
	MedleyDB	20	1h 20min
	RWC	24	2h 58min
Testing	ADC2004	12	4min
	MIREX 05	9	4min
	Medley DB	12	48min
	iKala	262	2h 6min

Table 1: The detailed descriptions of the datasets for training and testing the proposed framework ELH-SME.

Experiments

Datasets

We evaluate the proposed ELH-SME framework on several public datasets. Following (Yu 2024), we use the same training and testing data, and the datasets are listed in Table 1. The training data contains both labeled and unlabeled data. For the labeled data, we use 1000 popular music tracks from MIR-1K (Hsu and Jang 2010) and 30 popular music tracks from Medley DB (Bittner et al. 2014). For the unlabeled

[2]Notably, the representative feature in the memory bank keeps updating at each epoch.

Dataset Methods	ADC2004					MIREX 05				
	OA	RPA	RCA	VR	VFA	OA	RPA	RCA	VR	VFA
MSNet	70.1	71.3	73.2	75.6	21.3	81.7	76.7	76.9	83.6	18.6
MD+MR	71.2	72.4	73.8	73.1	24.7	80.8	77.2	77.8	81.3	24.8
Teacher-student	73.1	72.8	74.8	77.6	16.8	82.1	78.3	79.2	82.4	19.2
FTANet	72.4	73.8	75.2	77.3	24.9	84.4	79.7	80.0	83.8	5.1
HGNet	71.3	72.8	73.1	74.9	23.7	80.1	77.4	78.3	80.5	21.7
MCSSME	76.7	78.1	78.9	80.8	20.3	86.6	83.8	84.2	87.3	13.7
HKDSME	81.7	79.5	79.6	82.8	13.8	85.8	80.1	80.2	83.6	4.6
ELH-SME (ours)	**84.4**	**82.8**	**83.3**	**85.2**	**6.7**	**87.6**	**84.8**	**84.8**	**87.4**	**3.7**

Table 2: The performances of the proposed ELH-SME and baseline methods on the ADC2004 and MIREX 05 datasets, the values in the table are percentile.

Dataset Methods	Medley DB					iKala				
	OA	RPA	RCA	VR	VFA	OA	RPA	RCA	VR	VFA
MSNet	66.9	47.2	48.4	53.2	12.6	77.6	79.7	80.4	80.8	12.7
MD+MR	67.1	48.6	49.9	53.8	21.3	78.0	80.2	81.3	81.4	29.3
Teacher-student	68.1	49.0	49.6	58.3	29.7	76.7	76.4	78.2	79.1	36.9
FTANet	68.8	50.2	51.4	63.2	27.3	80.3	82.4	84.0	84.7	25.6
HGNet	65.3	45.4	46.2	51.7	24.9	78.7	80.0	80.6	81.5	24.9
MCSSME	73.4	56.8	57.4	64.2	16.2	84.7	85.8	86.2	88.3	9.2
HKDSME	72.2	60.5	61.4	66.2	13.5	83.2	84.1	84.2	87.1	11.3
ELH-SME (Ours)	**75.5**	**63.6**	**64.4**	**69.0**	**10.4**	**85.6**	**86.7**	**86.8**	**88.7**	**8.3**

Table 3: The performances of the proposed ELH-SME and baseline methods on the Medley DB and iKala datasets, the values in the table are percentile.

data we use 2000 unlabeled popular music tracks from the FMA dataset (Defferrard et al. 2017). Since some of our baseline methods contain an unsupervised domain adaptation (UDA) module, to conduct a fair comparison, we also choose 20 music tracks from MedleyDB and 24 music tracks from RWC for UDA setting. [3] For evaluation, we use four well-known testing datasets for this task: 12 tracks from ADC2004, 9 tracks from MIREX 05, 12 tracks from Medley DB and 262 tracks from iKala (Chan et al. 2015).

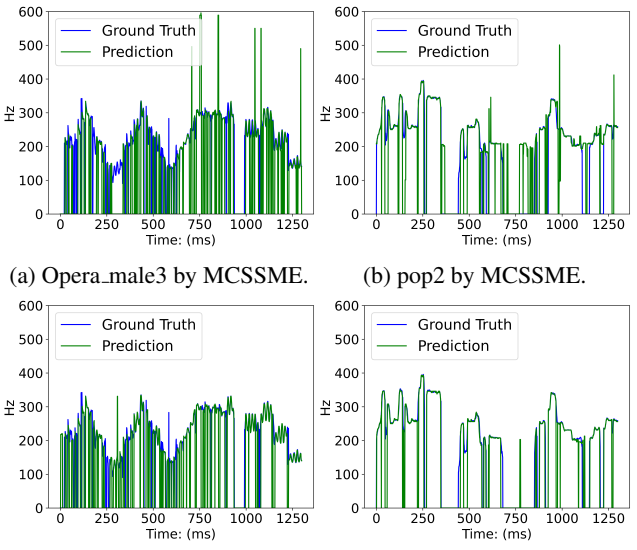
Experiment Setup

We followed the convention in previous studies to choose the following metrics for performance evaluation: overall accuracy (OA), raw pitch accuracy (RPA), raw chroma accuracy (RCA), voicing recall (VR) and voicing false alarm (VFA). We use mir eval library (Raffel et al. 2014) with the default setting to calculate the metrics. All metrics except for VFA, indicate better performance with higher scores. In the literature (Salamon et al. 2014), OA is often deemed more significant than other metrics.

Comparison with State-of-the-art Methods

We compare our framework with several state-of-the-art (SOTA) methods for singing melody extraction: (1) MSNet (Hsieh, Su, and Yang 2019), (2) MD+MR (Gao, You, and Chi 2020), (3) Teacher-student (Kum et al. 2020), (4) FTANet (Yu et al. 2021a), (5) HGNet (Yu, Chen, and Li 2022), (6) MCSSME (Yu 2024), (7) HKDSME (Yu et al.

[3]We only use these music tracks to train methods with UDA modules.



(a) Opera_male3 by MCSSME. (b) pop2 by MCSSME.

(c) Opera_male3 by our model. (d) pop2 by our model.

Figure 5: Visualization of singing melody extraction results using different models.

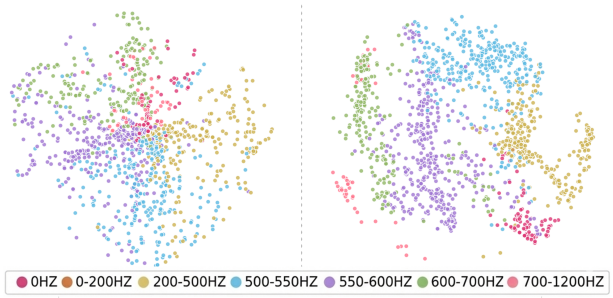


Figure 6: Visualization of the learned music representation via t-SNE. The left is the feature distributions using MCSSME, the right is the feature distributions using ELH-SME. Different colors denote various examples with different frequencies.

2024a). To demonstrate the effectiveness of our proposed method, we train the proposed framework ELH-SME and compare our method with other baseline methods. The quantitative results are shown in Table 2 and Table 3. It is observed the proposed ELH-SME achieves the best performance on four public testing sets in general. For comparison with other baselines, when focusing on OA, the proposed method outperforms MCSSME by 7.7% in ADC2004, by 1.1% in MIREX 05, by 2.1% in Medley DB and by 0.9% in iKala.

Case Study

To investigate what types of errors are solved by the proposed model, a case study is performed on two music tracks chosen from ADC2004 and MIREX 05 dataset. As depicted in Fig.5, we can observe that there are fewer errors in diagram (c) and (d) than in diagram (a) and (b). The result shows the proposed ELH-SME model can achieve better results in opera and popular music tracks than in MCSSME with fewer octave errors. Through the visualization, we can observe that the performance gains are from reducing octave errors via better utilization of unlabeled data.

To investigate the quality of music representation learned from our proposed ELH-SME, we visualize the learned representation via t-SNE. We use four opera tracks to perform t-SNE, as observed in Fig.6, the left is the feature distribution of the MCSSME. The right is the distribution of our ELH-SME. We can observe that representations are well clustered in the right. The visualized results demonstrates that ELH-SME enhances the representation learning in SSME task.

Ablation Study

To investigate the effectiveness of the key components in our framework, we conduct ablation studies and the quantitative results are presented in Table 4. First, DMA module is removed, and we only use the time-domain augmentation, the performances of OA decreased by 4.0% in ADC2004 and 4.5% in MIREX 05. The results justify the effectiveness of DMA module for SSME. We then remove the GCC module, the performances of OA decreased by 2.3% in ADC2004 and 1.7% in MIREX 05. The observation indicates that the use of global-class confidence helps to improve the performance of SSME. Next, we remove the CCA module, the performances of OA decreased by 1.7% in ADC2004 and 2.4% in MIREX 05. Overall, the key components of our framework ELH-SME are tightly incorporated and collaboratively devoted to promising results.

Dataset Methods	ADC2004			MIREX 05		
	OA	RPA	RCA	OA	RPA	RCA
w/o DMA	80.4	79.2	79.8	83.1	80.4	80.8
w/o GCC	82.1	81.8	81.9	85.9	82.8	83.5
w/o CCA	82.7	82.3	82.4	85.2	82.6	82.8
ELH-SME	**84.4**	**82.8**	**83.3**	**87.6**	**84.8**	**84.8**

Table 4: Results of Ablation Study on ADC2004 and MIREX 05 dataset.

Dataset Methods	ADC2004			MIREX 05		
	OA	RPA	RCA	OA	RPA	RCA
DDPM	81.2	79.4	79.5	85.9	80.6	80.6
Low freq.	81.1	79.4	79.5	85.7	80.5	80.6
Med. freq.	80.7	78.3	78.4	85.5	80.0	80.0
High freq.	80.2	77.9	78.0	84.8	79.8	79.8
DMA	**84.4**	**82.8**	**83.3**	**87.6**	**84.8**	**84.8**

Table 5: Effects of DMA on ADC2004 and MIREX 05 dataset.

To justify the assumption that we need to perform multi-bands data augmentations, we try to directly apply DDPM to the spectrogram. Meanwhile, we also investigate the separated augmentation on the unlabeled data. As shown in Tab.5, DMA outperforms all other four methods, and the results of DDPM are decreased by 3.2% in ADC2004 and 1.7% in MIREX 05. It is interesting to observe that the DDPM-based method performs similarly to using only low-frequency augmentation, highlighting the necessity of using a multi-bands data augmentation method.

Conclusion

In this paper, we have proposed a novel framework that better utilizes the unlabeled data, termed as ELH-SME. Specifically, we proposed a novel diffusion-based multi-bands augmentation (DMA) module to increase the amounts of training data. Then, to improve the utilization rate of the unlabeled data in SSME task, we proposed a global-class confidence (GCC) module to adaptively control the threshold for unlabeled samples. To further improve the utilization rate of unlabeled data, we also proposed a channel cross-attention (CCA) module that extracts channel-wise features from unlabeled musical audio, enhancing the representation capability of labeled data. We evaluated the performances on several well-known datasets. The experimental results demonstrate the effectiveness of our method.

References

Athiwaratkun, B.; Finzi, M.; Izmailov, P.; and Wilson, A. G. 2019. There Are Many Consistent Explanations of Unlabeled Data: Why You Should Average. In *Proc. ICLR*.

Azizi, S.; Kornblith, S.; Saharia, C.; Norouzi, M.; and Fleet, D. J. 2023. Synthetic Data from Diffusion Models Improves ImageNet Classification. *Transactions on Machine Learning Research*.

Berthelot, D.; Carlini, N.; Goodfellow, I. J.; Papernot, N.; Oliver, A.; and Raffel, C. 2019. MixMatch: A Holistic Approach to Semi-Supervised Learning. In *Proc. NeurIPS*, 5050–5060.

Bhosale, S.; Nag, S.; Kanojia, D.; Deng, J.; and Zhu, X. 2024. DiffSED: Sound Event Detection with Denoising Diffusion. In Wooldridge, M. J.; Dy, J. G.; and Natarajan, S., eds., *Proc. AAAI*, 792–800. AAAI Press.

Bittner, R. M.; McFee, B.; Salamon, J.; Li, P.; and Bello, J. P. 2017. Deep Salience Representations for F0 Estimation in Polyphonic Music. In *Proc. ISMIR*, 63–70.

Bittner, R. M.; Salamon, J.; Tierney, M.; Mauch, M.; Cannam, C.; and Bello, J. P. 2014. MedleyDB: A Multitrack Dataset for Annotation-Intensive MIR Research. In *Proc. ISMIR*, 155–160.

Brooks, T.; Holynski, A.; and Efros, A. A. 2023. Instructpix2pix: Learning to follow image editing instructions. In *Proc. CVPR*, 18392–18402.

Chan, T.; Yeh, T.; Fan, Z.; Chen, H.; Su, L.; Yang, Y.; and Jang, J. R. 2015. Vocal activity informed singing voice separation with the iKala dataset. In *Proc. ICASSP*, 718–722.

Chen, H.; Tao, R.; Fan, Y.; Wang, Y.; Wang, J.; Schiele, B.; Xie, X.; Raj, B.; and Savvides, M. 2023. SoftMatch: Addressing the Quantity-Quality Tradeoff in Semi-supervised Learning. In *Proc. ICLR*.

Chen, M.-T.; Li, B.-J.; and Chi, T.-S. 2019. CNN Based Two-stage Multi-resolution End-to-end Model for Singing Melody Extraction. In *Proc. ICASSP*, 1005–1009.

Cheng, D.; Ning, Y.; Wang, N.; Gao, X.; Yang, H.; Du, Y.; Han, B.; and Liu, T. 2022. Class-Dependent Label-Noise Learning with Cycle-Consistency Regularization. In *Proc. NeurIPS*.

Chou, H.; Chen, M.; and Chi, T. 2018. A Hybrid Neural Network Based on the Duplex Model of Pitch Perception for Singing Melody Extraction. In *Proc. ICASSP*, 381–385.

Defferrard, M.; Benzi, K.; Vandergheynst, P.; and Bresson, X. 2017. FMA: A Dataset for Music Analysis. In *Proc. ISMIR*, 316–323.

Du, X.; Zhu, B.; Kong, Q.; and Ma, Z. 2021. Singing Melody Extraction from Polyphonic Music based on Spectral Correlation Modeling. In *Proc. ICASSP*, 241–245.

Evans, Z.; Carr, C.; Taylor, J.; Hawley, S. H.; and Pons, J. 2024. Fast Timing-Conditioned Latent Audio Diffusion. In *Proc. ICML*.

Gao, P.; You, C.-Y.; and Chi, T.-S. 2020. A Multi-Dilation and Multi-Resolution Fully Convolutional Network for Singing Melody Extraction. In *Proc. ICASSP*, 551–555.

He, X.; Dong, K.; Cao, J.; Yu, S.; Li, W.; and Yu, Y. 2025. A Mamba-based Network for Semi-supervised Singing Melody Extraction Using Confidence Binary Regularization. In *Proc. ICASSP*, 1–5.

Hsieh, T.-H.; Su, L.; and Yang, Y.-H. 2019. A streamlined encoder/decoder architecture for melody extraction. In *Proc. ICASSP*, 156–160.

Hsu, C.; and Jang, J. R. 2010. On the Improvement of Singing Voice Separation for Monaural Recordings Using the MIR-1K Dataset. *IEEE Trans. Speech Audio Process.*, 18(2): 310–319.

Hu, Y.; Jing, J.; Li, F.; He, L.; Lin, L.; and Yang, W. 2025. A Singing Melody Extraction Network Via Self-Distillation and Multi-Level Supervision. In *Proc. ICASSP*, 1–5.

Islam, K.; Zaheer, M. Z.; Mahmood, A.; and Nandakumar, K. 2024. Diffusemix: Label-preserving data augmentation with diffusion models. In *Proc. CVPR*, 27621–27630.

Jeong, J.; and Shin, J. 2020. Consistency Regularization for Certified Robustness of Smoothed Classifiers. In *Proc. NeurIPS*.

Knees, P.; and Schedl, M. 2015. Music retrieval and recommendation: A tutorial overview. In *Proc. SIGIR*, 1133–1136.

Kum, S.; Lin, J.; Su, L.; and Nam, J. 2020. Semi-supervised learning using teacher-student models for vocal melody extraction. In *Proc. ISMIR*, 93–100.

Kum, S.; and Nam, J. 2019. Joint Detection and Classification of Singing Voice Melody Using Convolutional Recurrent Neural Networks. *Applied Sciences*, 9(7).

Kum, S.; Oh, C.; and Nam, J. 2016. Melody Extraction on Vocal Segments Using Multi-Column Deep Neural Networks. In *Proc. ISMIR*, 819–825.

Laine, S.; and Aila, T. 2017. Temporal Ensembling for Semi-Supervised Learning. In *Proc. ICLR*.

Li, S.; Zhang, Y.; Tang, F.; Ma, C.; Dong, W.; and Xu, C. 2024. Music Style Transfer with Time-Varying Inversion of Diffusion Models. In *Proc. AAAI*, 547–555.

Liu, J.; Dong, K.; Huang, Q.; Yu, S.; and Li, W. 2025. Ultra Lightweight Singing Melody Extraction via Combination of Convolution and MLP. In *Proc. ICASSP*, 1–5.

Lu, W. T.; Su, L.; et al. 2018. Vocal Melody Extraction with Semantic Segmentation and Audio-symbolic Domain Transfer Learning. In *Proc. ISMIR*, 521–528.

Ni, Y.; and Koniusz, P. 2023. NICE: NoIse-modulated Consistency rEgularization for Data-Efficient GANs. In *Proc. NeurIPS*.

Olsson, V.; Tranheden, W.; Pinto, J.; and Svensson, L. 2021. Classmix: Segmentation-based data augmentation for semi-supervised learning. In *Proc. CVPR*, 1369–1378.

Park, H.; and Yoo, C. D. 2017. Melody extraction and detection through LSTM-RNN with harmonic sum loss. In *Proc. ICASSP*, 2766–2770.

Raffel, C.; McFee, B.; Humphrey, E. J.; Salamon, J.; Nieto, O.; Liang, D.; Ellis, D. P.; and Raffel, C. C. 2014. mir_eval: A transparent implementation of common MIR metrics. In *Proc. ISMIR*.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. In *Proc. CVPR*, 10674–10685.

Saito, K.; Kim, D.; and Saenko, K. 2021. OpenMatch: Open-Set Semi-supervised Learning with Open-set Consistency Regularization. In *Proc. NeurIPS*, 25956–25967.

Salamon, J.; Gómez, E.; Ellis, D. P.; and Richard, G. 2014. Melody extraction from polyphonic music signals: Approaches, applications, and challenges. *IEEE Signal Processing Magazine*, 31(2): 118–134.

Serra, J.; Gómez, E.; and Herrera, P. 2010. Audio cover song identification and similarity: background, approaches, evaluation, and beyond. In *Proc. Advances in Music Information Retrieval*, 307–332. Springer.

Sohl-Dickstein, J.; Weiss, E. A.; Maheswaranathan, N.; and Ganguli, S. 2015. Deep Unsupervised Learning using Nonequilibrium Thermodynamics. In *Proc. ICML*, 2256–2265.

Su, L. 2018. Vocal melody extraction using patch-based CNN. In *Proc. ICASSP*, 371–375.

Tarvainen, A.; and Valpola, H. 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Proc. NeurIPS*, 1195–1204.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is All you Need. In *Proc. NeurIPS*, 5998–6008.

Wang, C.; and Jang, J. R. 2015. Improving Query-by-Singing/Humming by Combining Melody and Lyric Information. *IEEE/ACM Trans. Audio Speech Language Processing*, 23(4): 798–806.

Wang, S.; Kong, X.; Huang, H.; Wang, K.; and Hu, Y. 2025. HANet: A Harmonic Attention-Based Network for Singing Melody Extraction from Polyphonic Music. In *Proc. ICASSP*, 1–5.

Wang, Z.; Min, L.; and Xia, G. 2024. Whole-Song Hierarchical Generation of Symbolic Music Using Cascaded Diffusion Models. In *Proc. ICLR*.

Xie, Q.; Dai, Z.; Hovy, E. H.; Luong, T.; and Le, Q. 2020. Unsupervised Data Augmentation for Consistency Training. In *Proc. NeurIPS*, 6256–6268.

Xu, M.; Li, C.; Zhang, D.; Su, D.; Liang, W.; and Yu, D. 2024. Prompt-guided Precise Audio Editing with Diffusion Models. In *Proc. ICML*.

Yu, S. 2024. MCSSME: Multi-Task Contrastive Learning for Semi-supervised Singing Melody Extraction from Polyphonic Music. In *Proc. AAAI*, 365–373.

Yu, S.; Chen, X.; and Li, W. 2022. Hierarchical Graph-Based Neural Network for Singing Melody Extraction. In *Proc. ICASSP*, 626–630.

Yu, S.; He, X.; Chen, K.; and Yu, Y. 2024a. HKDSME: Heterogeneous Knowledge Distillation for Semi-supervised Singing Melody Extraction Using Harmonic Supervision. In *Proc. ACM Multimedia*, 545–553.

Yu, S.; He, X.; Dong, K.; and Yu, Y. 2025. DUDA: A Two-stage Decoupling Unsupervised Domain Adaptation Framework for Semi-supervised Singing Melody Extraction from Polyphonic Music. In *Proc. ACM Multimedia*, 8854–8862.

Yu, S.; He, X.; and Zhang, Y. 2024. RevNet: A Review Network with Group Aggregation Fusion for Singing Melody Extraction. In *Proc. ICME*, 1–6.

Yu, S.; Liu, J.; Yu, Y.; and Li, W. 2024b. A scalable sparse transformer model for singing melody extraction. In *Proc. ICASSP*, 1071–1075.

Yu, S.; Sun, X.; Yu, Y.; and Li, W. 2021a. Frequency-Temporal Attention Network for Singing Melody Extraction. In *Proc. ICASSP*, 251–255.

Yu, S.; Yu, Y.; Chen, X.; and Li, W. 2021b. HANME: Hierarchical Attention Network for Singing Melody Extraction. *IEEE Signal Process. Lett.*, 28: 1006–1010.

Yu, S.; Yu, Y.; Sun, X.; and Li, W. 2023. A neural harmonic-aware network with gated attentive fusion for singing melody extraction. *Neurocomputing*, 521: 160–171.

Zhang, B.; Wang, Y.; Hou, W.; Wu, H.; Wang, J.; Okumura, M.; and Shinozaki, T. 2021a. FlexMatch: Boosting Semi-Supervised Learning with Curriculum Pseudo Labeling. In *Proc. NeurIPS*, 18408–18419.

Zhang, S.; Qian, Z.; Huang, K.; Wang, Q.; Zhang, R.; and Yi, X. 2021b. Towards Better Robust Generalization with Shift Consistency Regularization. In *Proc. ICML*, volume 139, 12524–12534.