# LATENT FEATURE AUGMENTATION FOR CHORUS DETECTION

**Xingjian Du**[1] **Huidong Liang**[1] **Yuan Wan**[1] **Yuheng Lin**[1] **Ke Chen**[2] **Bilei Zhu**[1] **Zejun Ma**[1]
[1] ByteDance AI Lab, Shanghai, China
[2] University of California San Diego, San Diego, United States
duxingjian.real@bytedance.com

## ABSTRACT

In this paper, we introduce LA-Chorus, a chorus detection model based on **L**atent feature **A**ugmentation and ResNet-FPN architecture. We make three contributions. Firstly, we propose a method for implicitly augmenting chorus data in the latent space during the training stage. Compared to augmentations on audio surfaces such as time stretching and pitch shifting, latent augmentations indicate changes at a higher level in original audio, thereby increasing the diversity and sufficiency in training. Second, we apply Feature Pyramid Network (FPN) to generate additional embeddings from low dimension to high dimension, consequently achieving a multi-scale training paradigm. Lastly, we release Di-Chorus, a new diversified dataset of 13 genres and 14 languages for the community of music structure analysis. In conjunction with other public datasets, we conduct comprehensive experiments to evaluate the performance of our proposed method compared to other state-of-the-art models, where LA-Chorus outperforms other SOTAs by a considerable margin, meanwhile the proposed latent audio augmentation shows dominant advantages over traditional augmentation methods.

## 1. INTRODUCTION

Chorus detection, aiming for identifying the most "catchy" or "memorable" part of a song, is one of the fundamental tasks in music structure analysis (MSA) [1]. Chorus detection essentially helps better understand music compositions with computational modelling methods and has various applications, such as automatic chorus preview functions in music software that allow users to efficiently select songs from a large library according to their preference [2].

Currently, chorus detection models are based on deep neural network (DNN) architectures with a supervised MSA method, where annotations of different segments are used as target variables during the training stage [3–5]. The common approach is to regard chorus detection as a binary classification task, where each frame (or several frames) is assigned with a class label according to the corresponding segment annotation, and the model is trained to classify these labels [6].

To better perform this classification task, we identify two critical questions: (1) how to locate chorus with high precision as the resolution of feature maps decreases when model goes deeper, and (2) how to learn sufficient variations of chorus characteristics. The first issue requires incorporating chorus positional information into the latent representations learned by models, resembling the task of object detection in computer vision that aims to find the boundaries of target objects [7]. Nevertheless, as the network goes deeper, latent representations begin to lose positional meanings because of the reduced-sized feature maps (i.e. the resolution decreases) [8]. To address the problem of shrinking resolution in feature maps, previous chorus detection methods first used neural network architectures as backbones to generate audio embeddings, and then applied positional modifications on the networks. For example, [3] introduced a multi-task model that jointly detects chorus segments and their boundaries using convolutional neural network (CNN) to increase positioning precision, and [5] proposed a multi-scale CNN model that up-samples/down-samples the original audio features to better capture both global and local information. In this paper, we incorporate Feature Pyramid Networks (FPN) [8], a popular framework for object detection from computer vision, into a standard ResNet [9] as our model's backbone. It is designed specifically to tackle the problem of feature maps' decreasing resolutions by appending a network of reversed size order for each feature map with lateral connection, which will be discussed in Section 3.

The second issue, as learning sufficient chorus characteristics, can be addressed by using a diversified training dataset to feed the model such that it will generalize well at testing time. Nevertheless, despite the promising progress of emerging music annotations such as Isophonics [10], SALAMI [11], and Harmonics [12], the scarcity of labeled data (as they are costly to retrieve) and the deficiency of data diversity have always been challenging for music information retrieval (MIR) and other machine learning fields. To combat this problem, some traditional augmentation techniques have been proposed on the original inputs, such as rotating or flipping the images in computer vision [13], or time stretching [14] and pitch shifting [15] in audio signal processing. Recently, implicit augmentation methods, which focus on the augmentation in latent space, have shown remarkable performance over the pre-

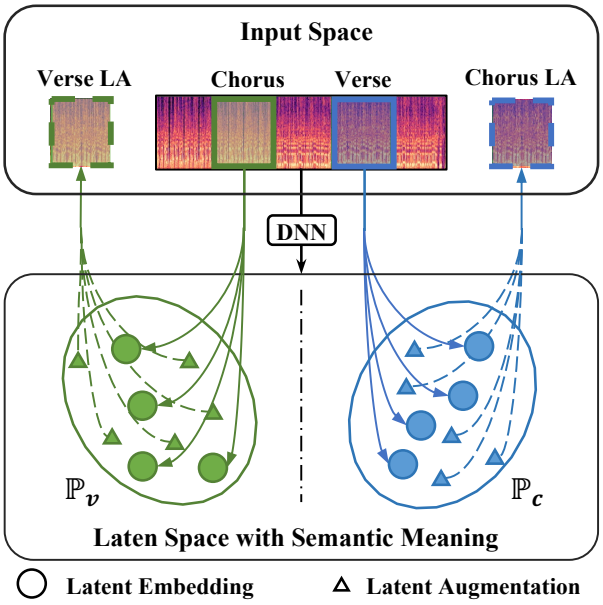


**Figure 1**. Illustration for implicit audio augmentation in MSA with two annotation types (verse and chorus) from a constant Q transformation (CQT) spectrogram excerpt of "Smooth Criminal" in *Isophonics* [10]. The spectrogram is first encoded into latent embeddings by frame, where chorus embedding and verse embedding follow latent distributions $\mathbb{P}_c$ and $\mathbb{P}_v$ respectively. Then latent augmentations are sampled around embeddings of verse/chorus segments, which correspond to augmented verse/chorus segments in the input space (represented by dashed lines, meaning they are NOT shown explicitly in the input space).

vious "shallow" methods in computer vision [16, 17]. The motivation is that latent representations may carry semantic meanings of original images (e.g. human age, gender, facial expressions) [18]. Such discoveries are also consistent with some works in audio signal processing, where latent audio representations have been found to capture distinctive audio features [19]. For example, [20] investigated the latent spaces of timbre and pitch of various instrument sounds encoded by a GM-VAE model; [21] introduced a Cycle-GAN based model for musical timbre transfer; and [22] disentangled pitch and rhythm representations to produce music analogies. According to these previous works, the latent features of audio correspond to semantic meaning of music and acoustic, such as timbre, rhythm pattern, etc. Therefore, augmentations in the latent space would correspond to changes of semantic features in the input audio segments, which leads to more variations than that of the augmentations in the shallow space. To our best knowledge, there is currently no application of latent augmentations in audio signal processing. In this paper, we illustrate this intuition by an MSA example in Figure 1.

Thus, we propose LA-Chorus, a supervised chorus detection model based on ResNet-FPN architecture that leverages implicit audio augmentations on latent features, which can better locate chorus positions and meanwhile enrich variations in training samples with semantic meanings. Moreover, since songs for most of the public datasets are not easy to retrieve, we further release a diversified collection of songs on YouTube for our MSA community, namely Di-Chorus, which contains 237 songs from 13 genres in 14 languages with annotations by experts. The rest of the paper is structured as follows: the next section introduces related works in chorus detection and latent augmentations, after which we discuss model structure and inference method. The experiment section presents LA-Chorus's performance on public datasets as well as Di-Chorus compared against other state-of-the-art models, followed by an ablation study showing the effectiveness of latent augmentations. Finally, the last section concludes our findings and contributions.

## 2. RELATED WORK

### 2.1 Chorus Detection

The origin of chorus detection tightly relates to thumbnailing, which aims to find a short preview (thumbnail) as a meaningful representation of a song [23]. Common approaches for thumbnailing includes evaluating the repeated sections of the audio waveform based on chroma transformation [24], selecting segments with the most repetition [25], and detecting significant change points with self-similarity matrix [26]. On the other hand, MSA assumes that songs contain different types of segments (e.g. chorus, verse, bridge, etc.) with certain structures [27]. Based on the assumption, many chorus detection algorithms in MSA took an unsupervised fashion in the early stage: [28] used heuristics to predict segment labels based on a restricted template for song structures; [29, 30] both applied Hidden Markov Model to derive different song sections; and [31] performed spectral clustering on the co-occurrence matrix generated from $k$-nearest neighbors.

With the advancement of deep learning in computer vision and natural language processing, DNN gradually makes its presence in MIR, among which ResNet [9], a CNN-based model with residual connections, becomes one of the most popular DNN architectures in recent MIR literature [4, 32]. At the same time, the emergence of labeled databases such as SALAMI [11] and Harmonix [12] make supervised learning gradually attractive for chorus detection, leading to the current paradigm of supervised chorus detection based on deep learning: [33] introduced a hybrid generative model with LSTM to directly predict segment labels; [3] proposed a multi-task method that jointly detects chorus segments and their boundaries; and [5] further proposed a multi-scale network with self-attention convolution to extract latent features of song segments, generating the current state-of-the-art results for chorus detection. In LA-Chorus, we will incorporate Feature Pyramid Network (FPN) [8], a top-down network that dedicates to the object positioning task, into ResNet as our model's backbone. The proposed framework is able to generate latent features with rich positional information and semantic meanings for later augmentation process, which will be discussed in detail in Section 3.

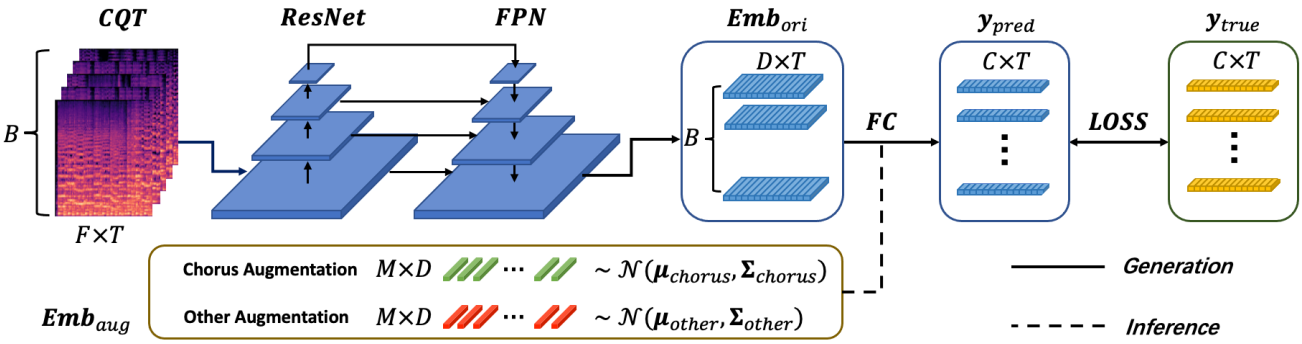


**Figure 2**. Model structure for LA-Chorus. The solid line represents the model generation processes in the forward pass, while the dashed line represents the inference (i.e. not explicitly generated in the forward pass).

### 2.2 Audio Data Augmentation

To enhance the diversity and variation in original audio features, traditional audio augmentation techniques such as time stretching, pitch shifting, noise perturbing and SpecAugment [34] are often implemented when transforming audio signals into spectrograms [14, 15]. In this paper, we take a different perspective: augmenting latent representations instead of original audio signals, motivated by the recent advancement of implicit augmentation methods in computer vision, represented by implicit semantic data augmentation (ISDA) and its variants [16, 17]. The ISDA model first used a backbone network to encode input images into latent space with semantic meaning, and then formulated a multi-variate Gaussian distribution for latent features in each class, which was estimated by their mean and covariance within the class via direct calculation. Then augmentations were sampled from the estimated distribution, and the model was later optimized by a novel cross-entropy loss tailored for latent augmentations. In this paper, we will demonstrate how this implicit augmentation method is utilized in audio to improve chorus detection.

## 3. PROPOSED METHOD

The structure of our proposed model is illustrated in Figure 2. We first use ResNet architecture to encode audio spectrograms into latent embeddings, Then, we apply the latent augmentation by sampling from estimated latent distributions for frames in the "chorus" class and the "non-chorus" (other) class respectively. Finally, the augmented representations are sent to a fully-connected layer to generate probability predictions. At learning and inference stages, we adopt a special cross-entropy loss that handles infinite number of latent augmentations via an upper bound, which saves the cost of sampling procedure.

### 3.1 ResNet-FPN

We first obtain the constant Q transform (CQT) spectrograms with $F$ frequency bins and $T$ frames in time domain after padding. Then a *ResNet-50* architecture is implemented as the embedding extractor $\mathcal{G}_\theta$ to extract latent embedding. Specifically, the ResNet consists of four stages that contains 3, 4, 6 and 3 residual CNN blocks respectively, where 64 convolution filters of $7 \times 7$ kernel size and a max-pooling layer of $3 \times 3$ kernel size are designed prior to these residual blocks to process the inputs.

With the size of each feature map reducing as CNN goes deeper, semantic information in deep audio features increases [8]. However, at the same time, the resolution of the feature map decreases and undermines the precision of chorus positioning. To solve this problem, we modify the backbone $\mathcal{G}_\theta$ by incorporating a Feature Pyramid Network (FPN) [8] into our ResNet architecture to construct latent features of high resolution from latent features with semantic information but of low resolution,as shown in Figure 2. The FPN design, different from the bottom-up ResNet part, takes a top-down approach that comprises four $1 \times 1$ convolutional layers that corresponds to four residual blocks of ResNet, which maps the low-dimensional outputs of ResNet to high-dimensional latent features with lateral connections.

As a result, the final latent representation $\mathbf{A}$ is of shape $T \times D$, where $T$ is the number of frames and $D$ is the number of latent dimensions. Because of the CNN and FPN designs in our backbone, latent embeddings not only contain both temporal and frequent information of the input audio, but also integrate feature maps of multiple resolutions to better locate chorus segments, further benefiting the latent augmentations later.

### 3.2 Latent Augmentations on Audio Features

To enrich variations, we apply latent augmentations on each representation $\mathbf{a}_i \in \mathbb{R}^D$ in the song, where $\mathbf{a}_i$ denotes the $i_{th}$ row in $\mathbf{A}$. Similar to other latent variable models such as VAE variants [35] and flow-based models [36], we make a fundamental assumption that latent features within the same class follow the same latent distribution. Specifically, a latent augmentation $\tilde{\mathbf{a}}_i \in \mathbb{R}^D$ for latent feature $\mathbf{a}_i$ follows a multi-variate Gaussian distribution $\mathcal{N}(\mathbf{a}_i, \mathbf{\Sigma}_{y_i})$, where $y_i$ indicates the label class ("chorus" or "other") for frame $i$, and $\mathbf{\Sigma}_{y_i} \in \mathbb{R}_+^{D \times D}$ represents the covariance matrix for class $y_i$. Then, we can sample from the distribution regarding to each $\mathbf{a}_i$ to get latent augmentations. In practice, a hyperparameter $\lambda > 0$ is imposed on the covariance matrix $\mathbf{\Sigma}_{y_i}$ to control the deviation of augmentations,

which leads to the final distribution of augmented latent representation $\tilde{\mathbf{a}}_i$:

$$\tilde{\mathbf{a}}_i \sim \mathcal{N}(\mathbf{a}_i\,, \lambda\boldsymbol{\Sigma}_{y_i}). \tag{1}$$

To estimate the covariance matrix $\boldsymbol{\Sigma}$ for different classes, we mathematically calculate the covariance matrix estimate $\hat{\boldsymbol{\Sigma}}_c$ for class $c$ ($c$ is either "chorus" or "other") within the dataset:

$$\hat{\boldsymbol{\Sigma}}_c = \frac{1}{N_c}\sum_{i=1}^{N_c}(\mathbf{a}_i - \bar{\mathbf{a}})(\mathbf{a}_i - \bar{\mathbf{a}})^T, \tag{2}$$

where $\bar{\mathbf{a}} = 1/N_c \sum_{i=1}^{N_c} \mathbf{a}_i$ is the mean of latent features in class $c$, and $N_c$ is the number of frames belong to class $c$. This covariance estimate is updated at each iteration after the generation of new latent features during training stage.

Finally, the augmented latent features $\tilde{\mathbf{A}}$ are sent to a fully connected layer with weight $\mathbf{W} \in \mathbb{R}^{D\times C}$ and bias $\mathbf{b} \in \mathbb{R}^C$ to generate the probability predictions, with $\tilde{\mathbf{a}}$ being row-vectors in $\tilde{\mathbf{A}}$:

$$\hat{\mathbf{y}} = \mathcal{F}_\phi(\tilde{\mathbf{A}}) = \tilde{\mathbf{A}}\mathbf{W} + \mathbf{b}. \tag{3}$$

In the next section, we will show an computationally efficient method for learning, which considers infinite augmentations but requires no explicit calculation of the augmented features $\tilde{\mathbf{A}}$.

### 3.3 Inference and Learning

Given a dataset of size $N$, a basic approach to formulate a loss function that treats each augmented latent features as a new sample point, and sample $M$ augmentations for each latent feature, which results in an augmented dataset of $N \times (M+1)$ samples. Then we use cross-entropy loss function to train the model. This method is effective when $M$ is relatively large, however, it is computationally expensive as we need to compute extra loss values for $N \times M$ augmentations.

Instead of computing the loss function with discrete augmentations, we adopt the loss function from [16] that incorporates latent augmentations from the continuous domain. For the final fully-connected layer, we denote $\mathbf{w}_c$ as the column vector in $\mathbf{W}$ and $b_c$ as the bias element in $b$ for class $c$, then the limit of the cross-entropy loss for $M$ augmentations when $M$ approaches to infinity equals:

$$\lim_{M\to\infty} -\frac{1}{N}\sum_{i=1}^{N}\frac{1}{M}\sum_{j=1}^{M}\log\Big(\frac{\exp(\mathbf{w}_{y_i}^T\mathbf{a}_i^{(j)} + b_{y_i})}{\sum_{c=1}^{C}\exp(\mathbf{w}_c^T\mathbf{a}_i^{(j)} + b_c)}\Big) \tag{4}$$

which is equivalent to calculating the expectation w.r.t. $\tilde{\mathbf{a}}_i$:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_{\tilde{\mathbf{a}}_i}\Big[\log\Big(\frac{\exp(\mathbf{w}_{y_i}^T\tilde{\mathbf{a}}_i + b_{y_i})}{\sum_{c=1}^{C}\exp(\mathbf{w}_c^T\tilde{\mathbf{a}}_i + b_c)}\Big)\Big] \tag{5}$$

$$= \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_{\tilde{\mathbf{a}}_i}\Big[\log\Big(\sum_{c=1}^{C}\exp(\tilde{\boldsymbol{\xi}})\Big)\Big] \tag{6}$$

where $\tilde{\boldsymbol{\xi}} = (\mathbf{w}_c - \mathbf{w}_{y_i})^T\tilde{\mathbf{a}}_i + (b_c - b_{y_i})$.

Since $\log()$ is a concave function, from Jensen's inequality, we can show:

$$\mathcal{L}_{upper} = \frac{1}{N}\sum_{i=1}^{N}\log\Big(\mathbb{E}_{\tilde{\mathbf{a}}_i}[\sum_{c=1}^{C}\exp(\tilde{\boldsymbol{\xi}})]\Big) \tag{7}$$

$$\geq \mathcal{L}. \tag{8}$$

As $\tilde{\mathbf{a}}_i$ follows $\mathcal{N}(\mathbf{a}_i\,, \lambda\boldsymbol{\Sigma}_{y_i})$ in Eqn. (1), and $\tilde{\boldsymbol{\xi}}$ is a linear transformation of $\tilde{\mathbf{a}}_i$, then $\tilde{\boldsymbol{\xi}}$ will also follow a Gaussian distribution:

$$\tilde{\boldsymbol{\xi}} \sim \mathcal{N}\big(\boldsymbol{\xi}\,, \Delta\big), \tag{9}$$

where $\boldsymbol{\xi} = (\mathbf{w}_c - \mathbf{w}_{y_i})^T\mathbf{a}_i + (b_c - b_{y_i})$ and $\Delta = \lambda(\mathbf{w}_c - \mathbf{w}_{y_i})^T\boldsymbol{\Sigma}_{y_i}(\mathbf{w}_c - \mathbf{w}_{y_i})$.

Given the moment generating equation $\mathbb{E}[\exp(tx)] = \exp(t\mu + \frac{1}{2}\sigma^2t^2)$ for $x \sim \mathcal{N}(\mu, \sigma^2)$, we can express Eqn. (7) as follows:

$$\mathcal{L}_{upper} = \frac{1}{N}\sum_{i=1}^{N}\log\Big(\sum_{c=1}^{C}\exp(\boldsymbol{\xi} + \frac{1}{2}\Delta)\Big) \tag{10}$$

$$= -\frac{1}{N}\sum_{i=1}^{N}\log\Big(\frac{\exp(\mathbf{w}_{y_i}^T\mathbf{a}_i + b_{y_i})}{\sum_{c=1}^{C}\exp(\mathbf{w}_c^T\mathbf{a}_i + b_c + \frac{1}{2}\Delta)}\Big), \tag{11}$$

which gives us a tractable upper-bound of Eqn. (6)

Therefore, we do not need to explicitly sample augmentations from its distribution in Eqn (1). Instead, only the covariance matrix $\boldsymbol{\Sigma}_c$ of latent features for each class requires update at each iteration, which speeds up the convergence compared to discrete estimation.

The algorithm for learning the embedding extractor $\mathcal{G}_\theta$ and the fully-connected layer $\mathcal{F}_\phi$ is demonstrated in Algorithm 1 below. Because the model is underfitted at beginning epochs, the hyperparameter $\lambda$ for controlling augmentation deviation is set to be $\lambda_0 \times \frac{\text{epoch}}{\text{total epoch}}$ to alleviate the impact of augmentation at the starting stage of training.

**Algorithm 1** Algorithm for training LA-Chorus

**Require**: Padded CQTs batches; ResNet extractor $\mathcal{G}_\theta$; A fully connected layer $\mathcal{F}_\phi$; Initial covariance matrix $\boldsymbol{\Sigma}_c$ for each class; Initial $\lambda_0$ for scaling augmentation

```
for epoch = 1, 2, ..., I do
    for batch = 1, 2, ..., K do
        Encode the CQT batch into latent features
        {a_i}_{i=1}^T via ResNet-FPN extractor G_θ
        Update G_θ and F_φ by computing:
        ∇_{θ,φ} L_upper from Eqn (11)
    end for
    Update Σ_c across all batches with Eqn. (2)
    λ ← λ_0 × epoch/I
end for
return G_θ and F_φ
```

## 4. EXPERIMENTS

In this section, we conduct comprehensive experiments to evaluate LA-Chorus's performance against other state-of-

| Dataset | #Songs | #Genres | #Lang. | #Quality |
|---|---|---|---|---|
| **Di-Chorus** | 237 | 13 | 14 | 3 |

**Table 1**. Key statistics of Di-Chorus.

the-art (SOTA) methods in chorus detection. We first introduce the experimental setups, where details of our newly released dataset Di-Chorus and hyperparameter settings in LA-Chorus are discussed. Then we present the results of chorus detection by LA-Chorus and other methods on popular public datasets, followed by an ablation study that demonstrates the effectiveness of different modules in our proposed method.

### 4.1 Experimental Setup

For a fair comparison purpose, we conduct our experiment under the same settings in [5] with a cross-dataset paradigm (i.e. testing on different datasets that are not used in training). Specifically, we use 890 songs that contain chorus segments in *Harmonix* [12], along with 38 songs of Michael Jackson and 83 songs of The Beatles in *Isophonics* set [10] for training and validation. At testing time, we use three public datasets and our released dataset Di-Chorus to evaluate our model and other methods. The public datasets used for testing are: 100 "Popular" songs from *RWC* [37, 38], 210 "Popular" songs (denoted as *SP*) and 198 "Live" songs (denoted as *SL*) from *SALAMI* [11], which are chosen in consistence with the testing sets in [5].

Our newly released dataset, Di-Chorus [1] (denoted as DC), contains 237 music annotations of songs on YouTube labeled by experts. Compared to previous datasets mentioned above, songs in Di-Chorus are more easy-to-access from the appended YouTube URLs, and are more diversified since it consists of musics tracks in 14 languages as opposed to other existing datasets that are mostly in English (e.g., Harmonics) or just two or three languages (e.g., RWC and SALAMI). In addition, we also include three different recording qualities to improve the variation within dataset: Studio, Live and Original Sound Track (OST) which contains non-music segments. The key statistics are summarized in Table 1 below.

To demonstrate the performance of our proposed model on the above datasets, we compare LA-Chorus against the following methods:

- **CNMF** [39]: an unsupervised matrix factorization method from *MSAF* [40].
- **SCluster** [31]: a spectral clustering method based on frame co-occurrence matrix from *MSAF* [40].
- **Highlighter** [41]: an CNN model that takes an unsupervised approach to detect emotional highlights as chorus segments.

[1] We provide some demos of Di-Chorus in the supplementary material. Di-Chorus will be made publicly available upon acceptance for retrieval.

| Models | AUC on Different Datasets | | | |
|---|---|---|---|---|
| | RWC | SP | SL | DC |
| *CNMF* | .526 | .543 | .478 | .488 |
| *SCluster* | .533 | .545 | .551 | .568 |
| *Highlighter* | .804 | .703 | .671 | .553 |
| *Multi2021* | .819 | .675 | .633 | - |
| *DeepChorus* | .842 | .780 | .765 | .811 |
| *LA-Chorus* | **.906** | **.887** | **.831** | **.872** |

**Table 2**. AUC results for chorus detection in various models.

| Models | F1-score on Different Datasets | | | |
|---|---|---|---|---|
| | RWC | SP | SL | DC |
| *CNMF* | .403 | .422 | .340 | .332 |
| *SCluster* | .427 | .448 | .392 | .603 |
| *Highlighter* | .407 | .303 | .251 | .283 |
| *Multi2021* | .643 | .473 | .380 | - |
| *DeepChorus* | .675 | .611 | .501 | .662 |
| *LA-Chorus* | **.728** | **.619** | **.526** | **.707** |

**Table 3**. F1-score results for chorus detection in various models.

- **Multi2021** [3]: a CNN model based on a multi-task learning objective that jointly predicts chorus segments and their boundaries.
- **DeepChorus** [5]: a CNN model based on multi-scale networks and self-attention, which is the current state-of-the-art method for chorus detection.

Then, we validate the prediction results by AUC score (Area Under Curve) and F1 score. To evaluate these two metrics, we first create a sequence of the song length from the original annotation, with each element indicating the class of the corresponding segment. Then we can calculate AUC and F1 score for each song independently and take the average over them as the final result.

For the training details, we resample the audio at 22050 Hz and use CQT as our input feature with 12 bins per octave, where Han windowing function is applied with a hop size of 512 for extraction. The model is trained for 100 epochs with a batch size of 32 and a learning rate of $10^{-4}$ with a cosine decay scheduler. The code is implemented in PyTorch and run at a Tesla-V100-SXM2-32GB GPU.

### 4.2 Chorus Detection

We retain the experiment results for the chosen SOTAs on RWC, SP and SL from [5], and test them on Di-Chorus with the default settings in their papers, as shown in Table 2 and Table 3 by AUC score and F1-score respectively. Note we do not test Multi2021 [3] on Di-Chorus, since their code is not open-sourced.

For AUC metric, LA-Chorus outperforms other SOTAs

on all datasets by a big margin. Compared to *DeepChorus* [5], which is considered as the current best method for chorus detection, our method improves the performance by over $0.06$ across all datasets. The performances on F1-score also exhibit a similar pattern, where LA-Chorus generates better predictions over other models with a considerable improvement on each dataset. In particular, our model performs exceptionally well on the widely used *RWC* dataset that reaches to an AUC score of $0.906$ and an F1-score $0.728$. We give most of the credits to the implicit augmentation design in our model, and we illustrate this perspective in the ablation study section.

### 4.3 Ablation Study

To analyze the effectiveness of FPN and latent augmentation, we test our LA-Chorus by separate modules: 1) ResNet backbone only, 2) ResNet with FPN, and 3) ResNet with latent augmentation (denoted as + *LA*). To demonstrate the efficacy of applying latent augmentations over traditional audio augmentation techniques in the input space, we further show the results of applying 4) time stretching (denoted as + *TS*) and 5) pitch shifting (denoted as + *PS*) to the ResNet backbone, with the results summarized in Table 4 for AUC and Table 5 for F1-score below (Note we do not apply TS or PS in our proposed method).

From the results, we can observe that by incorporating FPN, we improve the vanilla ResNet backbone by a remarkable increase of over $0.05$ on most of the datasets under both AUC and F1-score metrics, expect for *SALAMI-Live* where the result remains the same. Such findings indicate that FPN is an effective method to locate music segments by increasing the resolution of feature maps, which, without any augmentation, can already generate predictions that are comparative to *DeepChorus*.

On the other hand, when we apply implicit augmentations to latent features generated by ResNet (without FPN), significant improvements of over $0.10$ are witnessed for both AUC and F1 scores on most datasets. The results are even notably better than that of the *ResNet+FPN* combination who contains important positional information, implying the dominant role of latent augmentation in the strong performance of LA-Chorus. The results from *ResNet+TS* and *ResNet+PS* further corroborate the benefit of leveraging latent augmentations for chorus detection. Although there are some effects after adopting these two traditional augmentation methods, their improvements on the original model seem incremental compared to that of *ResNet+LA*. Instead, the implicit augmentation method outperforms traditional augmentation methods by a significant margin for both metrics on each dataset, which implies a clear advantage for adopting latent augmentations.

### 4.4 Discussion of Limitation

Despite of the prominent performance of LA-Chorus, we believe there are still some potential limitations for future explorations. First, further investigation is needed to verify that the latent augmentations are realistic to human when transforming them back to input domain. One possible way is to train a reversed model (such as a decoder or flow-based model) that reconstructs latent features to the original inputs. Second, the AUC and F1 metrics might not measure whether the output is overfragmented or underfragmented. It's needed to design metrics that are more perceptually relevant for chorus detection task. Finally, we only focus on detecting chorus segments in this paper, whereas in MSA, there are other annotation types (e.g. verse, bridge, etc.) to be modeled [4, 33]. We believe LA-Chorus only requires minor modifications in the class dimension of latent augmentations (i.e. augmenting chorus, verse, bridge, etc.) before being applied to predict other label types in music structure analysis.

| **Ablations** | AUC on Different Datasets | | | |
|---|---|---|---|---|
| | RWC | SP | SL | DC |
| *ResNet* | .801 | .773 | .767 | .751 |
| *ResNet + FPN* | .865 | .830 | .767 | .807 |
| *ResNet + LA* | .882 | .854 | .824 | .847 |
| *ResNet + TS* | .818 | .787 | .765 | .762 |
| *ResNet + PS* | .822 | .777 | .789 | .766 |
| *LA-Chorus* | .906 | .887 | .831 | .872 |

**Table 4**. AUC results for ablation study.

| **Ablations** | F1-score on Different Datasets | | | |
|---|---|---|---|---|
| | RWC | SP | SL | DC |
| *ResNet* | .592 | .415 | .418 | .588 |
| *ResNet + FPN* | .648 | .473 | .478 | .608 |
| *ResNet + LA* | .692 | .540 | .516 | .687 |
| *ResNet + TS* | .576 | .394 | .365 | .553 |
| *ResNet + PS* | .590 | .378 | .395 | .545 |
| *LA-Chorus* | .728 | .619 | .526 | .707 |

**Table 5**. F1-score results for ablation study.

## 5. CONCLUSION

In this paper, we introduced a novel chorus detection model based on ResNet-FPN architecture with latent augmentations on audio features. The proposed method, different from traditional augmentation algorithms focusing on the input space, augments audio features in the latent space to explore semantic changes in audio data. Besides, we released a new diversified dataset, Di-Chorus, with expert annotations, which contains songs with 13 genres in 14 languages and 3 qualities. Comprehensive experiments have been conducted on public datasets and Di-Chorus, where LA-Chorus shows superior performance against other methods. Lastly, the effectiveness of different modules in LA-Chorus are validated by an ablation study. In the future, we plan to investigate more details on the semantic changes of audio data via latent augmentations and the extensibility of LA-Chorus to other MIR tasks.

## 6. REFERENCES


[1] J. van Balen, J. A. Burgoyne, F. Wiering, R. C. Veltkamp *et al.*, "An analysis of chorus features in popular song," in *Proceedings of the 14th Society of Music Information Retrieval Conference (ISMIR)*, 2013.

[2] M. Goto, "A chorus-section detecting method for musical audio signals," in *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03).*, vol. 5. IEEE, 2003, pp. V–437.

[3] J.-C. Wang, J. B. Smith, J. Chen, X. Song, and Y. Wang, "Supervised chorus detection for popular music using convolutional neural network and multi-task learning," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 566–570.

[4] J.-C. Wang, J. B. Smith, W.-T. Lu, and X. Song, "Supervised metric learning for music structure feature," in *International Society for Music Information Retrieval Conference*, 2021.

[5] Q. He, X. Sun, Y. Yu, and W. Li, "Deepchorus: A hybrid model of multi-scale convolution and self-attention for chorus detection," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022.

[6] K. Ullrich, J. Schlüter, and T. Grill, "Boundary detection in music structure analysis using convolutional neural networks," in *Proceedings of the 15th International Society for Music Information Retrieval Conference (ISMIR)*, 2014.

[7] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 580–587.

[8] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2117–2125.

[9] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[10] M. Mauch, C. Cannam, M. Davies, S. Dixon, C. Harte, S. Kolozali, and D. Tidhar, "Omras2 metadata project 2009," in *In Late-breaking session at the 10th International Conference on Music Information Retrieval (ISMIR*, 2009.

[11] J. B. L. Smith, J. A. Burgoyne, I. Fujinaga, D. De Roure, and J. S. Downie, "Design and creation of a large-scale database of structural annotations." in *ISMIR*, vol. 11. Miami, FL, 2011, pp. 555–560.

[12] O. Nieto, M. McCallum, M. E. Davies, A. Robertson, A. M. Stark, and E. Egozy, "The harmonix set: Beats, downbeats, and functional segment annotations of western popular music." in *ISMIR*, 2019, pp. 565–572.

[13] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *Journal of big data*, vol. 6, no. 1, pp. 1–48, 2019.

[14] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio augmentation for speech recognition," in *Sixteenth annual conference of the international speech communication association*, 2015.

[15] J. Salamon and J. P. Bello, "Deep convolutional neural networks and data augmentation for environmental sound classification," *IEEE Signal processing letters*, vol. 24, no. 3, pp. 279–283, 2017.

[16] Y. Wang, X. Pan, S. Song, H. Zhang, G. Huang, and C. Wu, "Implicit semantic data augmentation for deep networks," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[17] S. Li, K. Gong, C. H. Liu, Y. Wang, F. Qiao, and X. Cheng, "Metasaug: Meta semantic augmentation for long-tailed visual recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 5212–5221.

[18] W. Liu, R. Li, M. Zheng, S. Karanam, Z. Wu, B. Bhanu, R. J. Radke, and O. Camps, "Towards visually explaining variational autoencoders," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[19] P. Agrawal and S. Ganapathy, "Interpretable representation learning for speech and audio signals based on relevance weighting," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2823–2836, 2020.

[20] Y.-J. Luo, K. Agres, and D. Herremans, "Learning disentangled representations of timbre and pitch for musical instrument sounds using gaussian mixture variational autoencoders," in *Proceedings of the 20th Society of Music Information Retrieval Conference (ISMIR)*, 2019, pp. 405–410.

[21] S. Huang, Q. Li, C. Anil, X. Bao, S. Oore, and R. B. Grosse, "Timbretron: A wavenet (cyclegan (cqt (audio))) pipeline for musical timbre transfer," in *International Conference on Learning Representations*, 2018.

[22] R. Yang, D. Wang, Z. Wang, T. Chen, J. Jiang, and G. Xia, "Deep music analogy via latent representation disentanglement," in *Proceedings of the 20th International Society for Music Information Retrieval Conference (ISMIR)*, 2019.

[23] W. Chai and B. Vercoe, "Music thumbnailing via structural analysis," in *Proceedings of the eleventh ACM international conference on Multimedia*, 2003, pp. 223–226.

[24] M. A. Bartsch and G. H. Wakefield, "To catch a chorus: Using chroma-based representations for audio thumbnailing," in *Proceedings of the 2001 IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics (Cat. No. 01TH8575)*. IEEE, 2001, pp. 15–18.

[25] M. Muller, N. Jiang, and P. Grosche, "A robust fitness measure for capturing repetitions in music recordings with applications to audio thumbnailing," *IEEE Transactions on audio, speech, and language processing*, vol. 21, no. 3, pp. 531–543, 2012.

[26] M. Cooper and J. Foote, "Automatic music summarization via similarity analysis." in *ISMIR*. Citeseer, 2002.

[27] J. Paulus, M. Müller, and A. Klapuri, "State of the art report: Audio-based music structure analysis." in *Ismir*. Utrecht, 2010, pp. 625–636.

[28] N. C. Maddage, C. Xu, M. S. Kankanhalli, and X. Shao, "Content-based music structure analysis with applications to music semantics understanding," in *Proceedings of the 12th annual ACM international conference on Multimedia*, 2004, pp. 112–119.

[29] G. Peeters, A. La Burthe, and X. Rodet, "Toward automatic music audio summary generation from signal analysis," in *ISMIR*, 2002, pp. 1–1.

[30] J. Paulus, "Improving markov model based music piece structure labelling with acoustic information." in *ISMIR*, 2010, pp. 303–308.

[31] B. McFee and D. P. Ellis, "Analyzing song structure with spectral clustering," in *Proceedings of the 15th Society of Music Information Retrieval Conference (ISMIR)*, 2014, pp. 405–410.

[32] X. Du, Z. Yu, B. Zhu, X. Chen, and Z. Ma, "Bytecover: Cover song identification via multi-loss training," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 551–555.

[33] G. Shibata, R. Nishikimi, and K. Yoshii, "Music structure analysis based on an lstm-hsmm hybrid model," in *International Society for Music Information Retrieval Conference*, 2020, pp. 15–22.

[34] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," *arXiv preprint arXiv:1904.08779*, 2019.

[35] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," in *International Conference on Learning Representations*, 2014.

[36] D. Rezende and S. Mohamed, "Variational inference with normalizing flows," in *International conference on machine learning*. PMLR, 2015, pp. 1530–1538.

[37] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, "Rwc music database: Popular, classical and jazz music databases." in *Ismir*, vol. 2, 2002, pp. 287–288.

[38] M. Goto *et al.*, "Aist annotation for the rwc music database." in *ISMIR*, 2006, pp. 359–360.

[39] O. Nieto and T. Jehan, "Convex non-negative matrix factorization for automatic music structure identification," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 236–240.

[40] O. Nieto and J. Bello, "Systematic exploration of computational music structure research," in *Proceedings of the 17th International Society for Music Information Retrieval Conference (ISMIR)*, 2016.

[41] Y.-S. Huang, S.-Y. Chou, and Y.-H. Yang, "Pop music highlighter: Marking the emotion keypoints," *Transactions of the International Society for Music Information Retrieval*, vol. 1, no. 1, 2018.