APPENDIX

A. CLASS ACTIVATION MAP AND t-SNE VISUALIZATION FOR TARGETING COMMON BIAS REGIONS

We expanded the experiment shown in Figure 1 of the original manuscript to demonstrate that our LCA mechanism can effectively direct the counterfactual attention branch to focus on the common biased regions, thereby encouraging the main branch to shift away from those regions. Similar to the experiment in Figure 1, we selected five artists from the *artist20* dataset (Ellis, 2007) for this experiment. In their recordings, we introduced white noise in the range of 3k to 5k Hz at varying data proportions. Specifically, the data of five artists - aerosmith, creedence_clearwater_revival, beatles, cure, and dave_matthews_band - had white noise added at proportions of 100%, 40%, 30%, 30%, and 30% respectively. Subsequently, we trained the CRNN_FGNL (Kuo et al., 2021) both with and without the LCA mechanism to visualize the class activation map (CAM) of five artists' noisy test data, utilizing Grad-CAM (Selvaraju et al., 2017). The hyperparameter settings for the experiment were the same as those for the evaluation protocols in Section 4.1. Compared to Figure A1 (b) without using the LCA mechanism, it's evident from Figure A1 (c) that after incorporating the LCA mechanism, CRNN_FGNL significantly shifts its attention away from the noise (bias) range. Such results confirm the effectiveness of our LCA mechanism in guiding the counterfactual attention branch to concentrate on common biased regions, subsequently prompting the main branch to divert its focus from these regions. Additionally, to confirm the effectiveness of our LCA mechanism in enhancing the model's ability to learn discriminative features, we employed t-SNE (van der Maaten & Hinton, 2008) for visualization and compared the feature distributions of CRNN_FGNL for these five artists under noisy data, both with and without the use of the LCA mechanism. Compared to Figure A2 (a), where the LCA mechanism is not used, Figure A2 (b) shows that with the LCA mechanism, CRNN_FGNL more effectively clusters the features of each artist in the embedding space and differentiates between artists, greatly reducing the impact of noise. These findings highlight the effectiveness of the proposed LCA mechanism.

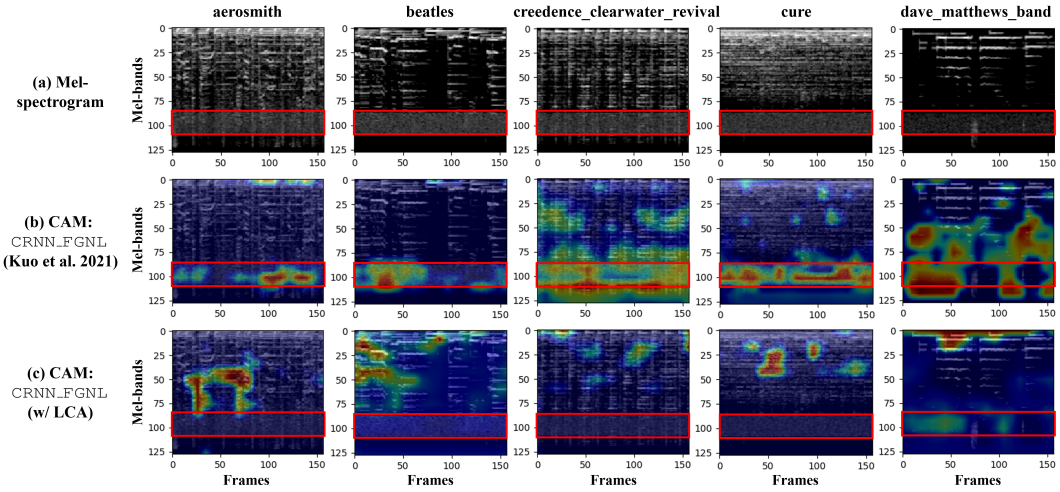
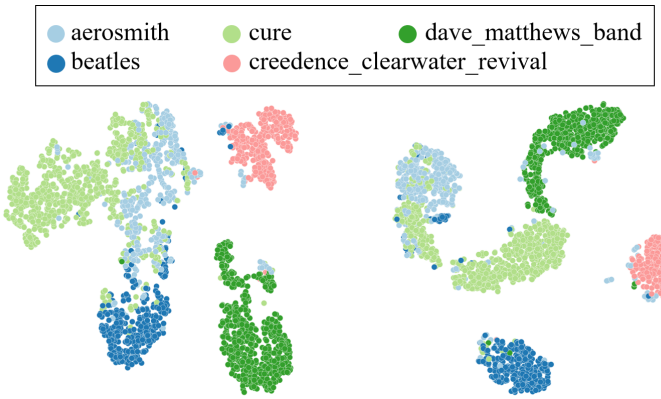


Figure A1: (a) The 5-sec Mel-spectrogram with white noise (highlighted in red frame) of songs from five different artists; (b) Class activation map (CAM) of CRNN_FGNL; (c) Class activation map (CAM) of our CRNN_FGNL (with LCA).



(a) t-SNE: CRNN_FGNL (Kuo et al. 2021)

(b) t-SNE: CRNN_FGNL (w/ LCA)

Figure A2: The t-SNE visualization of the features for (a) CRNN_FGNL and (b) our CRNN_FGNL (with LCA) under the 5-sec frame-level noisy test samples.

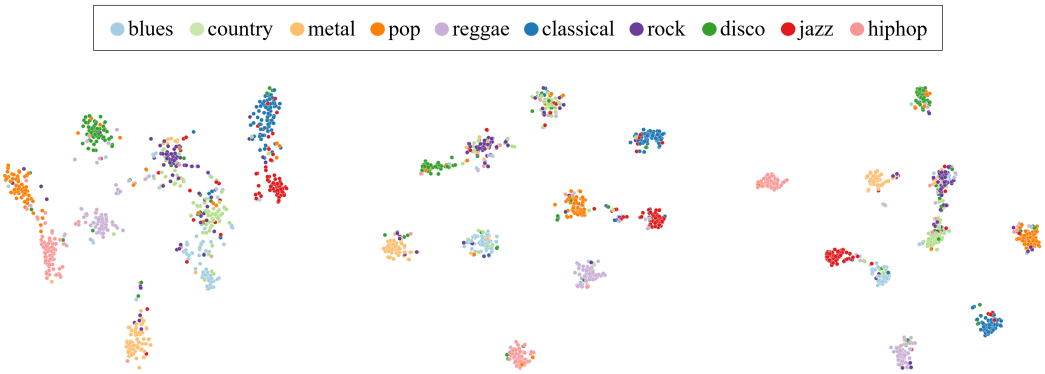
B. MUSIC GENRE CLASSIFICATION

To more conclusively demonstrate the proposed LCA mechanism's effectiveness, we expanded its evaluation to include music genre classification. For this, we employed the GTZAN dataset (Tzanetakis & Cook, 2002), which consists of ten genres, each with 100 audio files, 30 seconds each. We randomly divided the data from each genre into training, validation, and test sets, allocating 50%, 25%, and 25% to each set, respectively. Our experiments provided evaluation results for various versions of CRNN_FGNL, including original CRNN_FGNL (Kuo et al., 2021), CRNN_FGNL (with CAL) (Rao et al., 2021), and our CRNN_FGNL (with LCA), at both the frame and song levels. All models were trained following the methods outlined in the evaluation protocols of Section 4.1. The experiment, conducted using the original audio file setting, evaluated classification accuracy at both frame and song levels based on 10-second snippets.

The experimental results in Table A1 demonstrate that, compared to CAL (Rao et al., 2021), which uses random attention as counterfactual intervention, our LCA mechanism employing learnable counterfactual attention as the intervention for the main branch's attention during CRNN_FGNL traning leads to better performance improvements, both at the frame level and the song level. Even though our method (*i.e.,* CRNN_FGNL (with LCA)) only considers frame-level performance, its performance is already better than CAL (*i.e.,* CRNN_FGNL (with CAL)) at song level (including the use of a voting strategy). Furthermore, the results confirm that integrating the proposed LCA mechanism into the original CRNN_FGNL indeed significantly enhances the model's performance. The t-SNE visualization in Figure A3 clearly demonstrates the discriminative power of the features learned through our LCA mechanism. Additionally, since the LCA mechanism is only involved in the training phase, it does not add any computational complexity to the CRNN_FGNL during the testing phase. Overall, the experiment once again validated the effectiveness of the proposed LCA mechanism.

Table A1: Quantitative evaluation of music genre classification using different versions of CRNN_FGNL, including the original version, with CAL-enhanced version, and with our LCA-enhanced version, at both frame and song levels.

	Original Audio File		
Models	**Frame Level**	**Song Level**	**#Para.**
(Classification accuracy %)	10s	10s	(M)
CRNN_FGNL (Kuo et al., 2021)	72.6	77.2	0.58
CRNN_FGNL (w/ CAL) (Rao et al., 2021)	74.3	78.4	0.58
CRNN_FGNL (w/ LCA) Ours	**78.9**	**82.3**	0.58



(a) t-SNE : CRNN_FGNL (b) t-SNE : CRNN_FGNL (w/ CAL) (c) t-SNE : CRNN_FGNL (w/ LCA)

Figure A3: The t-SNE visualization of the features for (a) CRNN_FGNL, (b) CRNN_FGNL (with CAL), and (c) our CRNN_FGNL (with LCA) under the 10-sec frame-level music genre test samples.

C. HYPERPARAMETER SETTING: WEIGHT OF EACH LOSS FUNCTION IN EQUATION (1)

In the SID task, the weight of each loss function is as follows:

- Original audio file setting: λ_{ce}^{main}= 1.0, λ_{ce}^{effect}= 1.0, λ_{ce}^{cf}= 0.3, λ_{ent}^{cf}= 0.25, λ_{1}^{att}= 1.0, λ_{ent}^{main}= 0.2
- Vocal only setting: λ_{ce}^{main}= 1.0, λ_{ce}^{effect}= 1.0, λ_{ce}^{cf}= 0.8, λ_{ent}^{cf}= 0.025, λ_{1}^{att}= 1.0, λ_{ent}^{main}= 0.02

For the newly implemented music genre classification task, we have set the weight of each loss function to be the same as those used in the original audio file setting of the SID task.

REFERENCES

Daniel P. W. Ellis. Classifying music audio with timbral and chroma features. In *International Society for Music Information Retrieval Conference (ISMIR)*, 2007.

I-Yuan Kuo, Wen-Li Wei, and Jen-Chun Lin. Positions, channels, and layers: Fully generalized non-local network for singer identification. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2021.

Yongming Rao, Guangyi Chen, Jiwen Lu, and Jie Zhou. Counterfactual attention learning for fine-grained visual categorization and re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

George Tzanetakis and Perry R. Cook. Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing*, 10(5):293–302, 2002.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008.