# Supplementary Material: BirdSoundsDenoising: Deep Visual Audio Denoising for Bird Sounds

Youshan Zhang
Yeshiva University, NYC, NY
youshan.zhang@yu.edu

Jialu Li
Cornell University, Ithaca, NY
jl4284@cornell.edu

## 1. Noise STFT images

In Fig. 1, we use raw STFT images minus the ground truth clean bird sound areas to get the noise signal STFT images. We could observe that most of these noise areas are concentrated in the center of the image.

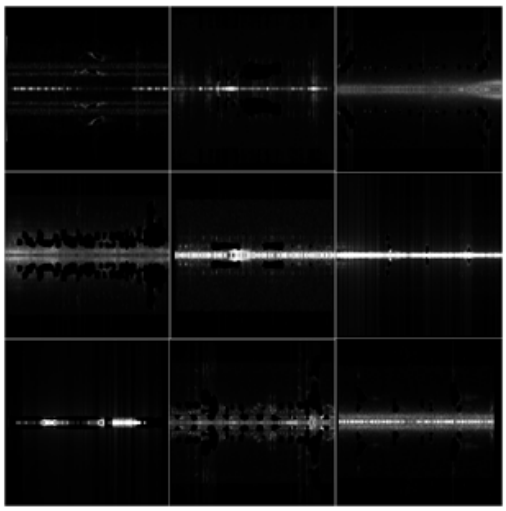

Figure 1: Nine examples of different kinds of noise signal areas in STFT images.

## 2. Audio ImageMask tool

Fig. 2 shows the interface of our proposed audio ImageMask tool. This software is developed based on MATLAB 2021a. The top menu bar supports the following functions (from left to right).

1. open the audio image
2. go to the previous image
3. go to the next image
4. accept the labeled image
5. the eraser
6. zoom in
7. zoom out
8. the pan to move the image
9. undo
10. redo
11. use the pentagon to label the image, and this is the key function to label the image.

There are also many other supported functions. The order of the following explanation is from top to bottom and from left to right.

1. The edit text box: it will show the index number of current audio image. "1" is the default initialization number. We can also directly change this number to jump to another image.
2. In the switch "Mask" radio botton box: it supports the switch between "Ground Truth" and "Predicted label". The "Ground Truth" is the labeled mask, and "Predicted label" is the predicted mask from other segmentation models. Hence, we can compare the differences between human-labeled masks and predicted labels from deep models.
3. "255_Bird_Audio": it means that the mask is for bird audio, the saved mask for clean bird signal areas is 255, and its background is 0 (while the clean areas will be 1 during the model training).
4. The edit text box and slider bar can change the transparency of the labeled mask. As shown in Fig. 3, the mask uses the transparency of 0.5. If this threshold is higher, the mask color will be deeper. If it is lower, the mask color will be lighter.
5. "CleanMask": it will remove all masks for re-labeling.

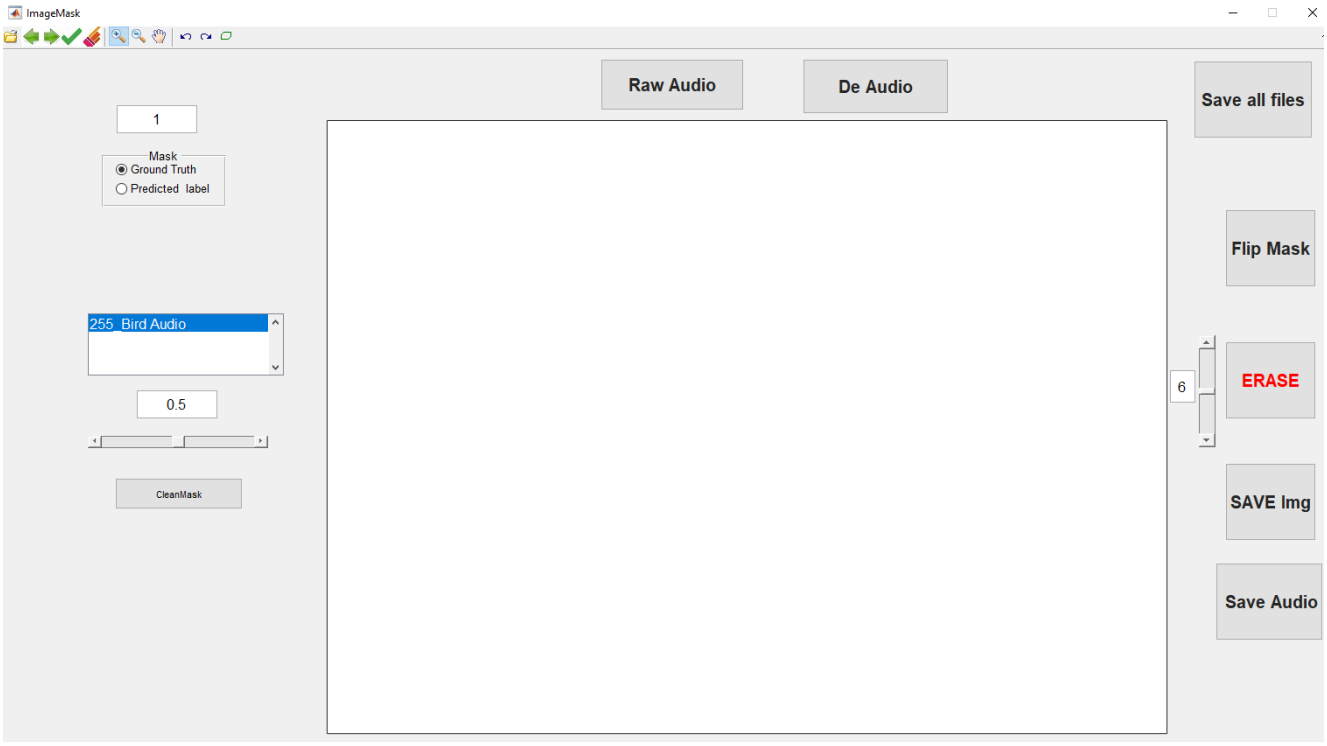


Figure 2: The interface of our proposed audio ImageMask tool.

6. Static text area: it will show some texts after finishing some steps. E.g., "Image saved!" will be displayed after finishing the "SAVE Img" step.

7. "Raw Audio": it will play the raw bird audio with noise.

8. "De Audio": it can play the denoised bird audio.

9. Image information area: it will demonstrate the current image number, total image number, and image name, e.g., "59/145:XC3087.png.

10. "Save all files": this button can save all human-verified labeled files. It will save all accepted files in a folder named "Accepted". In the folder, it contains another four sub-folders: original audio, denoised audio, audio images, and audio masks.

11. "Flip Mask": the audio is a symmetrical image, we only need to label half of the image, then we can click "Flip Mask" to replicate the rest part of the mask.

12. The slider bar and "ERASE": it supports the eraser function to remove small and wrongly labeled areas. The slider bar can adjust the size of the eraser.

13. "Save Img": it can save all labeled masks

14. "Save Audio": it can save the denoised audio. Then, we can listen to it using the "De Audio" function.

15. Shortcut functions: the left and right arrows in the keyboard have the same function as "go to the previous image" and "go to the next image". "Ctrl + q" has the same function of "using pentagon to label image" as in the menu bar. After that, we could click the boundary of the image. We can also use zoom in and zoom out to check the details of the boundaries. Finally, we could use double-click to finish the image labeling and get the red areas of the labeled mask as shown in Fig. 3.

## 3. Supported audio file explanation

We also list many supported denoised audios for the results in the main paper.

1. **All denoised audio comparisons.** In the folder "01_all_results_comparisons", it contains 11 audios. "XC234266_left" is the audio file name. "raw_noise_audio" is the original noisy bird sound audio. "labeled_denoised_audio" is our labeled denoised audio. From file number 3 to 11, these are the denoised results of nine methods. We also present the audio signals in Fig. 4. Our proposed DVAD model with DeepLabV3 as the segmentation model achieves better performance

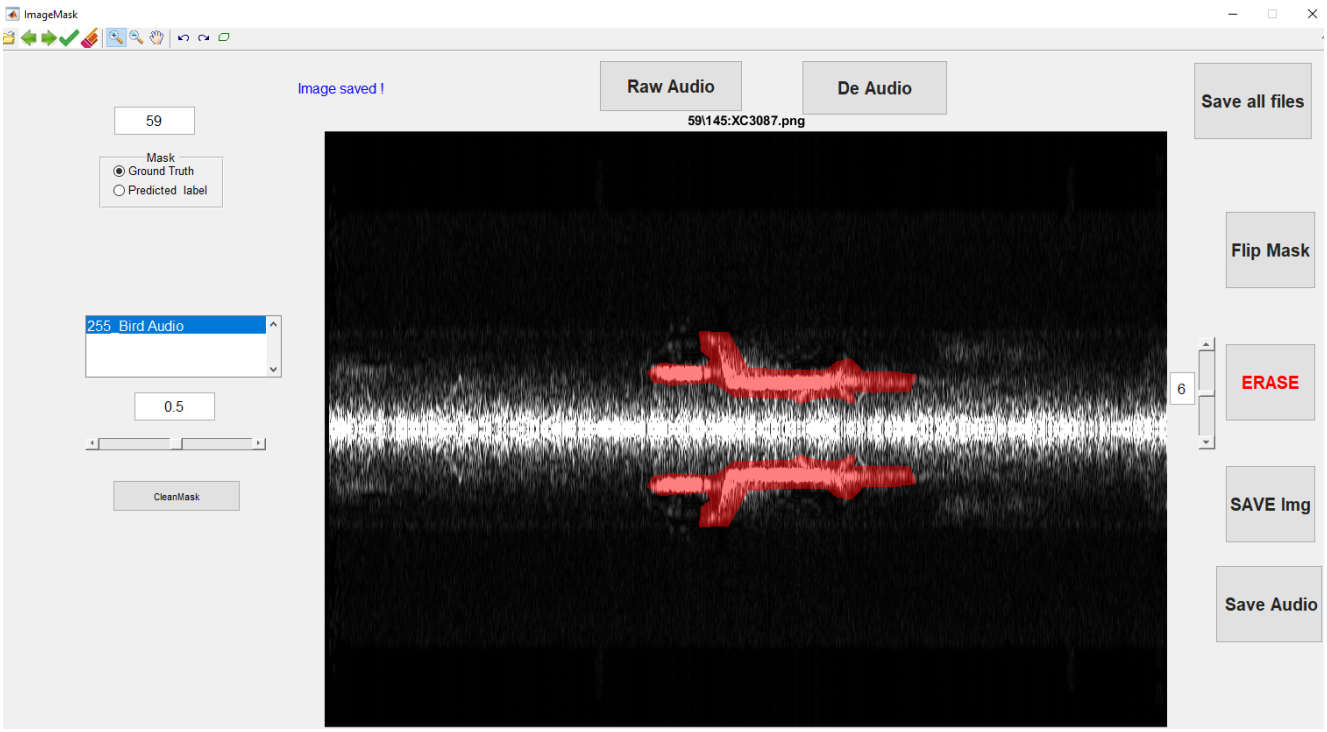


Figure 3: An example of labeled audio image mask. The red color is the labeled mask. On the top, "59/145": 59 is the current audio image number, and 145 is the total number of images. "XC3087.png" is the audio image name. "Image saved!" will be displayed after clicking the "SAVE Img" button.

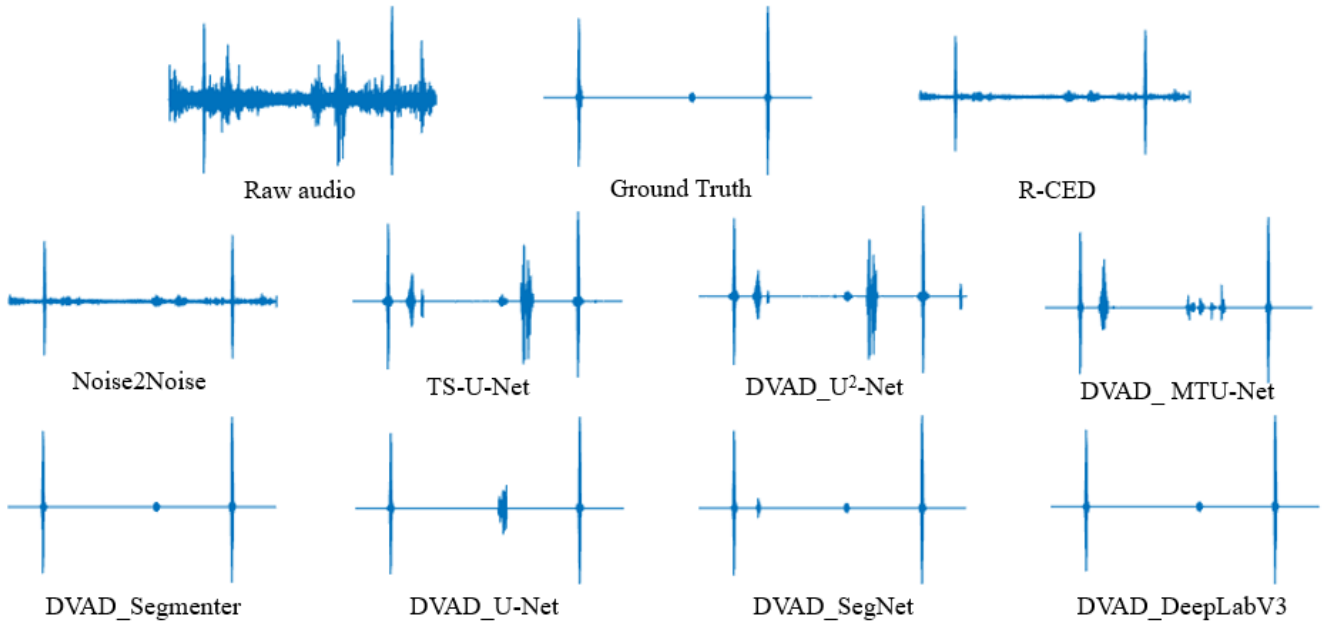


Figure 4: All denoising results comparisons. Raw audio is the original noise audio. Ground truth is the labeled mask.

than most models. We can still hear the sound of flies in most audios. "DVAD_Segmenter" model surprisingly perform well in this example.

2. **Speech_denoising.** In the "02_speech_denoising" folder, it contains denoising results of a noisy human speech audio. "1_human_speech_raw_audio" is the raw

noisy audio and “2_human_speech_denoised_audio” is the denoised audio using our DVAD model as presented in Fig. 8 in the main paper.

3. **Audio separation.** In the “03_audio_separation” folder, it contains four audios. “1_XC603581_left_raw_audio” is the noise audio. “2_XC603581_left_denoised_audio” is the denoised audio. “3_XC603581_left_denoised_separated_bird_1” is the separated bird sound one, while “4_XC603581_left_denoised_separated_bird_2” is the separated bird sound two as shown in Fig. 9 in the main paper.

4. **Audio enhancement.** In the “04_audio_enhancement”, it contains three audios. “1_XC585071_left_raw_audio” is the original noise audio. “2_XC585071_left_denoised_audio” is the denoised audio. “3_XC585071_left_denoised_enhanced” is the enhanced audio with 200 times of audio signal as stated in Fig. 10 in the main paper.