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

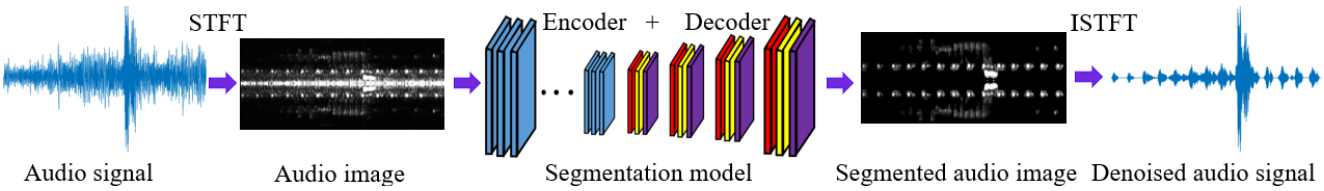


Figure 1: The overall progress of our proposed deep visual audio denoising model (DVAD).

Abstract

Audio denoising has been explored for decades using both traditional and deep learning-based methods. However, these methods are still limited to either manually added artificial noise or lower denoised audio quality. To overcome these challenges, we collect a large-scale natural noise bird sound dataset. We are the first to transfer the audio denoising problem into an image segmentation problem and propose a deep visual audio denoising (DVAD) model. With a total of 14,120 audio images, we develop an audio ImageMask tool and propose to use a few-shot generalization strategy to label these images. Extensive experimental results demonstrate that the proposed model achieves state-of-the-art performance. We also show that our method can be easily generalized to speech denoising, audio separation, audio enhancement, and noise estimation.

1. Introduction

With the development of technology, audio signals have been increasingly used as main sources of information transmission [17], such as teleconferences [14], the speech-to-text function in social media [32], the lung [22] and heart [15] sounds for disease diagnosis, instrument solo identification [11], and hearing aid [37, 27, 33], etc. Therefore, it is important to maintain the quality of signal transmission and retain as much useful information as possible. However, due to existing noises in the actual environment, the transmission of audio signals, including speech and other signals that we intend to collect, are inevitably affected, resulting in the poor quality and intelligibility of audio signals. Audio denoising can significantly increase audio quality and contribute to a better outcome of information transition.

Audio denoising has been a popular research area in recent years and different methods have been applied to reduce noise and separate audio, including traditional statistics approaches [5, 28, 12, 19] and deep learning approaches [27, 26, 2, 24, 16]. While there are several difficulties encountered across these models. In this paper, we specifically use samples from the natural environment, which presents more challenges to the proposed research models.

Why natural audio denoising is difficult?

Firstly, the most common difficulty encountered is the limited sources for training. Deep learning-based models require both clean and noisy audio samples for training. However, in reality, audio signals come with noises that cannot be separated to produce desired training samples [14]. Secondly, most noisy audio samples used for model training are artificially compiled, such as white gaussian noise (WGN) [30, 39], which is composed differently from natural noise. In addition, we could still observe the clean signal patterns in the artificial noise audio, while it is difficult to observe the clean signal patterns in real noise audio as shown in Fig. 1 (leftmost and rightmost signal). Therefore, the denoising performance of the training models might not perform as well in the real setting compared to experiments.

These two challenges are commonly encountered in the audio denoising field, and we address them using a deep visual audio denoising model (DVAD). In this paper, we first collect audio samples that are directly acquired from the natural environment. The proposed model can process more complex and natural noises compared to previous models. We offer three principal contributions:

- We present a benchmark bird sounds denoising dataset with the goal of advancing the state-of-the-art in audio denoising under natural noise background.
- To the best of our knowledge, we are the first to transfer

audio denoising into an image segmentation problem. By removing the noise area in the audio image, we can realize the purpose of audio denoising.

- We develop an audio ImageMask tool to label the collected dataset and apply a few-shot generalization strategy to accelerate the data label process. We also demonstrate that our model can be easily extended to speech denoising, audio separation, audio enhancement, and noise estimation.

2. Related Work

Audio denoising has been widely explored, and many methods have evolved from traditional methods of estimating the difference between noise and clean audio statistics [37], to the adoption of deep learning methods [3].

Traditional methods for audio denoising can be dated back to the 1970s. Boll [5] proposed a noise suppression algorithm for spectral subtraction using the spectral noise bias calculated in a non-speech environment. Another statistical method proposed in [28] is a more comprehensive algorithm, combining the concept of A Priori Signal-to-Noise Ratio (SNR) with earlier typical audio enhancement schemes such as Wiener filtering [6, 18], spectral subtraction, or Maximum Likelihood estimates. In the realm of the frequency-domain algorithm, minimum mean square error (MMSE) based approaches is a mainstream approach besides Wiener filtering. Hansen et al. [12] proposed an auditory masking threshold enhancement method by applying Generalized MMSE estimator in an auditory enhancement scheme. In [19], the MMSE estimator is used to enhance the performance of short-time spectral coefficients by estimating discrete Fourier transform (DFT) coefficients of both noisy and clean speech. One major problem is that the performance of traditional methods for noise separation and reduction will be degraded with the presence of natural noises, which are largely different from artificial noises applied in the experiments.

Wavelet transformation methods are developed to overcome the difficulty of studying signals with low SNR and are reported with better performance than filtering methods. Zhao et al. [41] used an improved threshold denoising method, overcame the discontinuity in hard-threshold denoising, and reduced the permanent bias in soft-threshold denoising. Srivastava et al. [30] developed a wavelet denoising approach based on wavelet shrinkage, allowing for the analysis of low SNR signals. Pouyani et al. [22] proposed an adaptive method based on discrete wavelet transform and artificial neural network to filtrate lung sound signals in a noisy environment. Kui et al. [15] also combined the wavelet algorithm with CNNs to classify the log mel-frequency spectral coefficients features from the heart sound signal with higher accuracy. These combined methods outperformed single wavelet transformation methods.

Deep learning methods are later introduced to the audio denoising field, complementing the disadvantages of traditional methods and demonstrating a stronger ability to learn data and characteristics with a few samples [37]. The deep neural network (DNN)-based audio enhancement algorithms have shown great potential in their ability to capture data features with complicated nonlinear functions [16]. Xu et al. [38] introduced a deep learning model for automatic speech denoising to detect silent intervals and better capture the noise pattern with the time-varying feature. Saleem et al. [27] used the deep learning-based approach for audio enhancement accompanying complex noises. An ideal binary mask (IBM) is used during the training and testing, and the trained DNNs are used to estimate IBM. Xu et al. [39] proposed a DNN-based supervised method to enhance audio by finding a mapping function between noisy and clean audio samples. A large mixture of noisy dataset is used during the training and other techniques, including global variance equalization and the dropout and noise-aware training strategies. Saleem et al. [26] also developed a supervised DNN-based single channel audio enhancement algorithm and applied less aggressive Wiener filtering as an additional DNN layer. Vuong et al. [34] described a modulation-domain loss function for a deep learning-based audio enhancement approach, applying additional Learnable spectro-temporal receptive fields to enhance the objective prediction of audio quality and intelligibility.

Yet a problem in speech denoising application of DNN is that sometimes, it is difficult for models to track a target speaker in multiple training speakers, which means that the DNNs are not easy to handle long-term contexts [33, 16]. Therefore, deep learning approaches, such as convolutional neural network (CNN)-based and recurrent neural network (RNN)-based models, are explored. Alamdari et al. [2] applied a fully convolutional neural network (FCN) for audio denoising with only noisy samples, and the study displayed the superiority of the new model compared to the traditional supervised approaches. Germain et al. [10] trained an FCN using a deep feature loss, trained for acoustic environment detection and domestic audio tagging. The research showed that this new approach is particularly useful for audio with the most intrusive background noise. Kong et al. [14] proposed an audio enhancement method with pre-trained audio neural networks using weakly labeled data with only audio tags of audio clips and applied a convolutional U-Net to predict the waveform of individual anchor segments selected by PANNs. Raj et al. [24] proposed a multilayered CNN-based auto-CODEC for audio signal denoising, using the mel-frequency cepstral coefficients, providing good encoding and high security. Abouzid et al. [1] combined the convolutional and denoising autoencoders into convolutional denoising autoencoders for the suppression of noise and compression of audio data.

On top of single-type deep learning methods, Tan et

al. [33] proposed a recurrent convolutional network by incorporating a convolutional encoder-decoder and long short-term memory (LSTM) into the convolutional recurrent neural network (CRN) architecture to address real-time audio enhancement. This method outperformed an existing LSTM-based model with fewer trainable parameters. Gao et al. [8], and [9] respectively applied a progressive learning framework for DNN-based and LSTM-based audio enhancement to improve model performance and reduce complexity. Li et al. [16] combined the progressive learning framework with a causal CRN to further reduce the trainable parameters and improve audio quality and intelligibility. This proposed method produced a close performance to the CRN.

Many deep learning approaches are implemented in the time-frequency domain, using short-time Fourier transform (STFT) and inverse short-time Fourier transform (ISTFT) [36]. Some methods address audio enhancement via time-domain algorithms, viewing audio enhancement as a filtering problem [40]. Yu et al. [40] proposed a DNN-based Kalman filter algorithm for audio enhancement. The DNN is used for estimating the linear prediction coefficients in the KF. Sonning et al. [29] investigated the performance of a time-domain network for speech denoising, addressing the original inability of STFT/ISTFT-based time-frequency approaches to capture short-time changes and discovered its usefulness in a real-time setting. Wang et al. [36] proposed a two-stage transformer neural network for end-to-end audio denoising in the time domain, including an encoder, a two-stage transformer module , a masking module and a decoder. Their model outperformed many time- or frequency-domain models with less complexity.

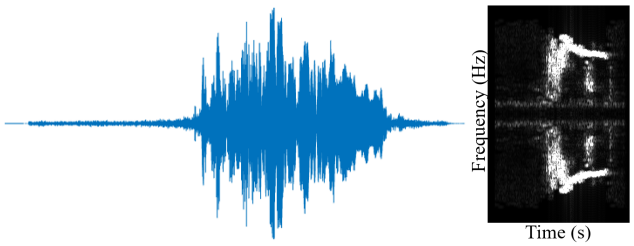
3. Methods

3.1. Problem

Given a noisy audio signal $\{x_t\}_{t=1}^{T}$, we aim to extract the clean audio $\{y_t\}_{t=1}^{T}$ by learning a mapping $\mathcal{M}$. The goal of audio denoising is to minimize the approximation error between the denoised audio $\{\mathcal{M}(x_t)\}_{t=1}^{T}$ and clean audio $\{y_t\}_{t=1}^{T}$. In our DVAD model, we convert audio denosing to an image segmentation problem. Given the audio images $\mathcal{I} = \{I^i\}_{i=1}^{n}$ based on audio signals $X = \{x^i\}_{i=1}^{n}$ and its ground truth labeled masks $M = \{m^i\}_{i=1}^{n}$, we propose to minimize the error between prediction of any image segmentation model $F(I)$ and M.

3.2. Motivation

Although some existing deep audio denoising models utilized magnitude images of audio signals, they only filtered out some regions of the image to realize the purpose of denoising. The details of these images are less explored. Our DVAD model delves into the audio image to find different patterns between noise and clean signal areas. As shown in Fig. 2b, we can find that there are significant differences between the noise and clean signal areas. Therefore, we can achieve the purpose of audio denoising if we can segment the clean signal areas. We further treat the audio denoising as an image segmentation problem.



(a) Audio signal (b) Audio image

Figure 2: The conversion from audio signal (a) to audio image (b) by taking the absolute value of STFT. (b) is a symmetrical image, and most noise areas are concentrated in the center of an image (the light white horizontal lines). The informative and clean signal (bird sound) majorly lies in the bright white patterns. Please refer to supplementary material for more noise signal areas.

3.3. Preliminary

3.3.1 Short-Time Fourier Transform (STFT)

STFT is used to analyze how the frequency content of a nonstationary signal changes over time.

$$STFT_x(t, f) = \int_{-\infty}^{\infty} x(t)\omega(t-\tau)e^{-j2\pi ft}dt \qquad (1)$$

where $STFT_x(t, f)$ is the coefficient of STFT. STFT is a function of time (t) and frequency (f), and it shows how frequency f of the signal $x(t)$ changes with time t. ω is a window function, τ is a short time and j is the square root of -1. In our model, we aim to convert the signal to frequency domain and get the raw images for each bird sound.

3.3.2 Inverse Short-Time Fourier Transform (ISTFT)

The STFT is invertible, $i.e.$, the original signal can be reconstructed from the transform by the inverse STFT. It is defined as:

$$\tilde{x_t} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} STFT_x(t', f')\omega(t-t')e^{-j2\pi f't'}dt'\,df'. \qquad (2)$$

In our model, we aim to reconstruct the bird sound based on the segmented bird sound image.

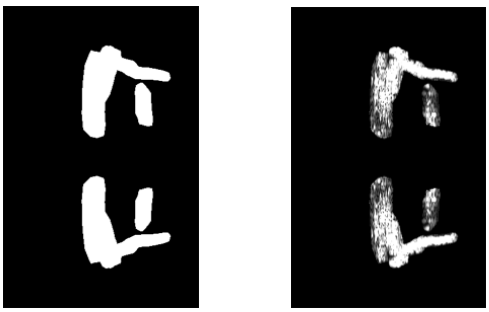
3.4. Methodology

To form the audio denoising problem as an image segmentation problem, we first need to represent the audio in image format. After performing STFT using Eq. (1), let

$S = STFT_x(t, f)$, we can define the audio image (I) in the following equation,

$$I = abs(S), \tag{3}$$

where abs takes the absolute value from the complex frequency domain S. As shown in Fig. 2, we convert a one-second bird sound audio to its audio image. We observe that the patterns of noise areas and clean sound areas are distinguishable. If we segment the clean sound areas in the audio image, then we could remove the noise from the frequency domain S. As shown in Fig. 3b, we can apply ISTFT in Eq. (2) after getting the denoised audio image to reconstruct the denoised audio signal. Therefore, we can convert the audio denoising problem into an image segmentation problem.



(a) Mask image (b) Segmented image

Figure 3: The mask (a) and segmented audio image (b) for the signal in Fig. 2. (a) is the mask of clean signal areas in Fig. 2b and (b) is the segmented audio image by removing noise areas.

To realize the image segmentation task, we need to train a segmentation model F to segment the clean audio signal area. In our DVAD model, we train the segmentation model using dice loss as follows.

$$Dice\ loss = 1 - 2 \times \frac{m \cap \tilde{m}}{m + \tilde{m}}, \tag{4}$$

where m is the ground truth mask and $\tilde{m} = F(I)$ is the predicted mask of the segmentation model given the input image I. In the mask, we denote the clean audio areas using 1 and represent the noise areas using 0. After training F, we can predict the segmented mask of any audio image. Next, we aim to reconstruct the denoised audio.

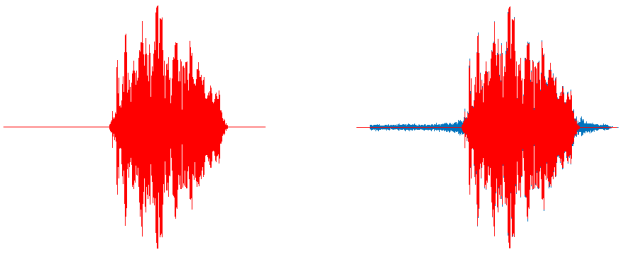
In Eq. (2), we can recover the original $x(t)$ given the key input frequency domain S. To remove the noise audio, we need to filter out noise areas in S given the predicted mask from the segmentation model. We define the new frequent domain S' in the following equation:

$$S' = S, \quad \text{and} \quad S'[\tilde{m} < 1] = 0, \tag{5}$$

where $S'[\tilde{m} < 1] = 0$ aims to replace all noise area with 0 to realize the purpose of removing noise areas. Then, we can apply ISTFT to reconstruct the denoised audio as follows.

$$\tilde{x_t} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} S'\omega(t - t')e^{-j2\pi f' t'} dt'\, df'. \tag{6}$$

As shown in Fig. 4a, we can get the denoised audio after removing the noise audio using Eq. (5). We also show the overlapping of original signal with denoised signal in Fig. 4b. The rest part of blue signal (noise areas) is removed from the red denoised signal.



(a) Denosied audio (b) Overlapping signals

Figure 4: The denoised audio using our DVAD model (a) and overlapping of original signal (blue color) and denoised audio signal (red color).

3.5. DVAD overall algorithm

Considering all steps in Sec. 3.4, the scheme of our proposed DVAD model is shown in Fig. 1 and the overall algorithm is presented in Alg. 1.

Algorithm 1 Deep Visual Audio Denoising (DVAD). $B(\cdot)$ denotes the mini-batch training sets, I is the number of iterations.

1: **Input:** Audio signals $X = \{x^i\}_{i=1}^n$ and labeled mask images $M = \{m^i\}_{i=1}^n$, where n is the total number of audios.
2: **Output:** Denoised audio signals
3: Generate audio images $\mathcal{I} = \{I^i\}_{i=1}^n$ using Eq. (3)
4: **for** $iter = 1$ **to** I **do**
5: Derive $B(\mathcal{I})$ and $B(M)$ sampled from $\mathcal{I}$ and M
6: Optimize any segmentation model F using Eq. (4)
7: **end for**
8: Get the clean frequency domain using Eq. (5)
9: Output the denoised audio signals using Eq. (6)

4. Datasets

4.1. Data Collection

Our data is collected from the xeno-canto website, which is a public website to share bird sounds from around the world [1]. We first collect 15,300 bird sounds from one second to fifteen seconds. Unlike many audio denoising datasets,

[1] https://xeno-canto.org/explore

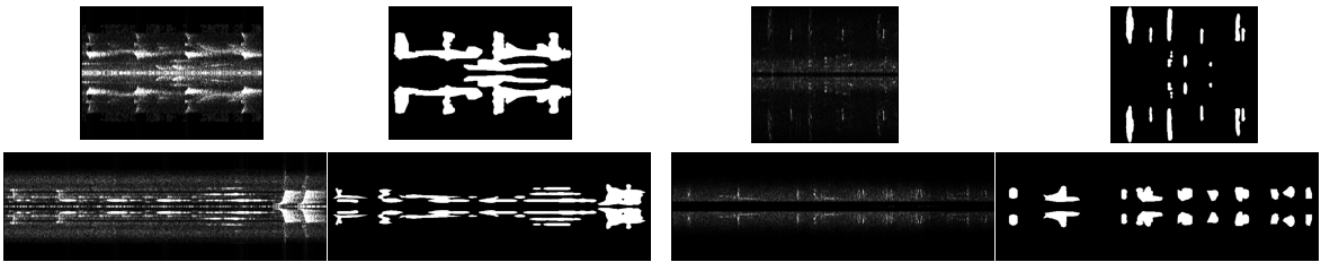


Figure 5: Four sample images and masks. In each sample, the left is the audio image, and the right is its labeled mask. The width of each sample depends on the length of the audio. Longer audio will produce a wider audio image.

which have manually added artificial noise, our collected bird sounds contain natural noises, including wind, waterfall, rain, etc. Then, we apply STFT in Eq. (1) to convert bird sounds to the frequency domain and get the audio images using Eq. (3). Some audios have two soundtracks (left and right soundtracks). Hence we get more images than collected audios. Since we have converted audio denoising to an image segmentation problem, we need to provide masks for audio images to optimize any image segmentation models. Therefore, our next task is to label these audio images.

4.2. Data labeling

Image mask labeling is time-consuming, tedious, and expensive. However, to train any machine learning algorithms, we have to provide enough labeled datasets to achieve a good performance. Given that there is no specific software for audio image labeling, we also developed an audio ImageMask tool. The ImageMask software has three key functions. (1). It can open an audio image and label it to create a mask and show the overlapping between the raw audio image and labeled mask. (2). We can save created masks and denoised audios. The software also supports human verification. All accepted denoised audio will be saved in an 'Accepted' folder. The folder contains another four subfolders: original audio, denoised audio, audio images, and audio masks. (3). We can also compare the ground truth masks with predicted masks from any segmentation models to validate the performance of models. More details of our developed audio ImageMask tool can be found in the supplementary material.

Few-shot generalization

Although we developed a specific audio denoising software, it takes around 5 minutes to label one audio image. We have more than 15,000 audio images, and it is still time-consuming to label all images. To accelerate the labeling process, we propose to utilize the few-shot generation to first predict coarse masks for audio images. Then we can verify and update these coarse masks to get better masks.

Few-shot learning aims to learn a robust model based on a few labeled samples, then improve the performance of new datasets. To ease the process of image labeling from scratch, we first manually labeled 100 audio images as training and 40 images as test. We select DeepLabV3 [7] as the segmentation model and train the DeepLabV3 model using these 140 labeled images to get a basic model F. We could then predict the coarse mask via $F(I)$. Given any unlabeled audio image I^i, we can get all predicted coarse masks as $\{F(I^i)\}_{i=1}^n$. Finally, these coarse masks can be further modified using our developed audio ImageMask tool. After using the proposed few-shot generalization strategy, the whole dataset is labeled by four experts in one month.

4.3. BirdSoundsDenoising dataset

After finishing the data labeling process using Sec. 4.2, we can save all accepted labeled audio to create the BirdSoundsDenoising dataset (note that some low-quality audio will be removed during the labeling process). The BirdSoundsDenoising dataset contains 14,120 audio images and has three folders: training, validation, and test. In each folder, another four sub-folders are included: raw audios, denoised audios, images, and masks. Tab. 1 shows the statistics of each folder. As shown in Fig. 5, we list four audio images and their labeled masks. These audio images are varied, and clean signal areas are also different. Longer audio can produce a longer image. In the experiments, we will fine-tune models using training and validation datasets and report results on the test dataset.

Table 1: Statistics on BirdSoundsDenoising dataset

Datasets	Training	Validation	Test
Number of samples	10,000	1,400	2,720

4.4. Dataset creation details

In STFT, converting audio signals to audio images, we first use 128-point Hamming as the window function, the number of overlapped samples $= 64$, the number of DFT points $= 1024$, to change bird sounds to frequency domain S, and save all bird audio images using Eq. (3). In few-shot generalization, we utilize DeepLabV3 as a basic segmentation model to generate coarse audio image masks. During the training of 140 images, we set batch size = 16, training iteration $I = 100$, learning rate = 0.0001 with an Adam opti-

mizer on a RTX A6000 GPU. The input image size of the DeepLabV3 model is $[512 \times 512 \times 3]$. We excluded audio images with no bird sound or extreme noisy background.

5. Experiments

To evaluate the performance of our proposed DVAD model, we test it on our created BirdSoundsDenoising dataset. We first train six different state-of-the-art segmentation models to demonstrate the effects of different segmented masks on audio denoising performance. These six selected segmentation models have an encoder-decoder architecture. The encoder aims to extract important features from images (e.g., edge), and the decoder learns how to map these low-resolution features to the prediction at the pixel level.

1. SegNet [4]: the encoder network utilizes the layers from VGG16, and the decoder network is followed by a pixel-wise classification layer. The decoder of SegNet can upsample its lower-resolution input feature map(s) using pooling indices computed in the maxpooling step of the corresponding encoder to perform non-linear up-sampling.

2. U-Net [25]: it has a structure called dilated convolutions and removes the pooling layer structure. The U-Net architecture comprises a contracting path to capture context, and a symmetric expanding path that enables precise localization.

3. DeepLabV3 [7]: it uses dilated convolutions and a fully connected conditional random field to implement the atrous spatial pyramid pooling (ASPP), which is an atrous version of SPP and can account for different object scales and improve the accuracy.

4. U^2-Net [23]: it is able to capture more contextual information from different scales with the mixture of receptive fields of different sizes in ReSidual U-blocks.

5. Segmenter [31]: it utilizes the Vision Transformer (ViT) as the encoder to encode all image patches. A point-wise linear decoder is applied to patch encodings. It also includes a decoder with a mask transformer to further improves the performance.

6. MTU-Net [35]: it proposes a novel transformer module named Mixed Transformer Module (MTM). It calculates self affinities efficiently by a Local-Global Gaussian-Weighted Self-Attention (LGG-SA). It also mines inter-connections between data samples using External Attention (EA).

5.1. Implementation details

The training settings of six different segmentation methods are the same as Sec. 4.4 except that we use batch size of 12 for MTUNet and 32 for Segmenter[2]. To show the superiority of the DVAD model, we also compare the proposed model with three audio denoising methods [21, 13, 20].

5.2. Results

5.2.1 Evaluation metrics

We use three metrics ($F1, IoU$ and $Dice$) to evaluate the performance of image segmentation as follows.

$$F1 = \frac{TP}{TP + \frac{1}{2}(FP + FN)}, \\ IoU = \frac{m \cap \tilde{m}}{m + \tilde{m}} \quad Dice = 2 \times \frac{m \cap \tilde{m}}{m + \tilde{m}}, \tag{7}$$

where TP is the number of true positives, FP is the number of false positives, and FN is the number of false negatives. For audio denoising, we use signal-to-distortion ratio (SDR) to evaluate our DVAD model. The higher these four metrics, the better the segmentation model is.

$$SDR = 10 \log_{10} \frac{||m||^2}{||\tilde{m} - m||^2} \tag{8}$$

5.3. Performance comparisons

We first show the comparisons of six different segmentation models in Fig. 6. The segmented masks of the DeepLabV3 model are better than the other five models. Similarly, we could observe that DeepLabV3 has the highest $F1, IoU$, and $Dice$ scores in Tab. 2. Therefore, we can infer that the DeepLabV3 model is the best segmentation model for our BirdSoundsDenoising among all six segmentation models. In addition, we also reported the mean SDR of all bird sounds in both validation and test datasets. As shown in Tab. 2, the SDR score of our DVAD model with DeepLabV3 as the segmentation model achieves the highest value. Notably, the performance level of three audio denoising methods (R-CED, Noise2Noise, and TS-U-Net) is relatively lower than all other segmentation models. The comparisons of raw bird audio, true labeled denoised audio, and denoised audio from other models are shown in Fig. 7. The denoised signal of DVAD (with DeepLabV3) is also closer to the labeled denoised signal. Therefore, our DVAD architecture is effective in improving the audio denoising performance.

6. Discussion

We compared six different state-of-the-art segmentation models and three deep audio denoising methods. One obvious strength of our model is its better performance than other methods. Especially the different variants of our DVAD

[2]BirdSoundsDenoising dataset and code are available at https://github.com/YoushanZhang/BirdSoundsDenoising

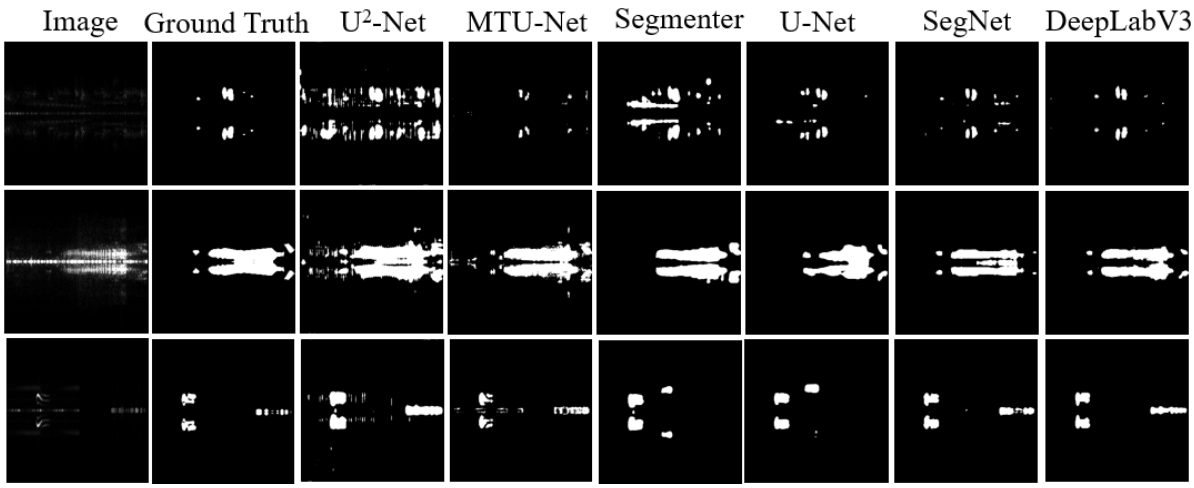


Figure 6: Segmentation results comparisons. Leftmost column is the original audio image. Ground truth is the labeled mask.

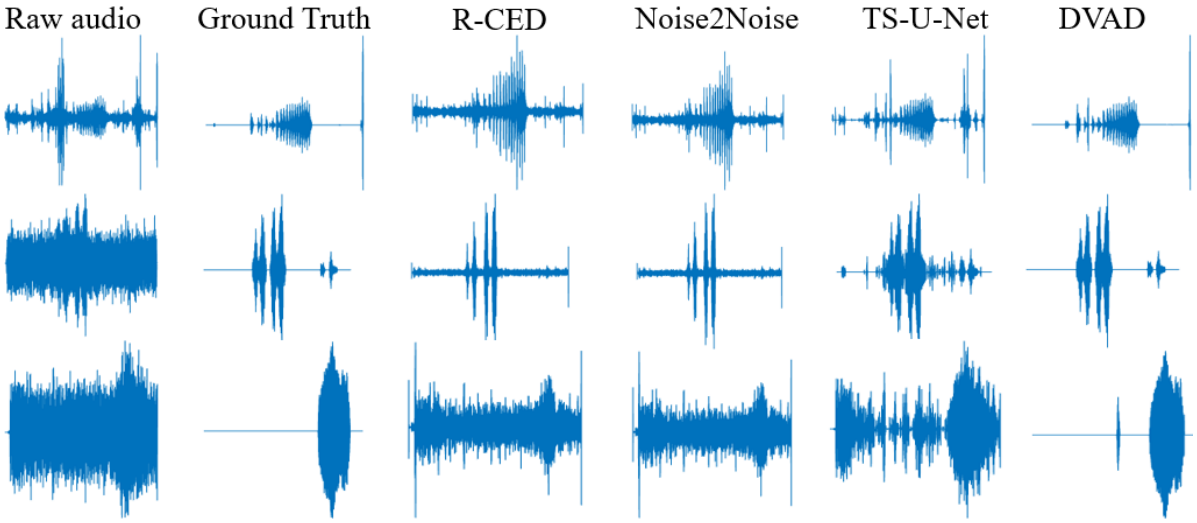


Figure 7: Denoising results comparisons. Raw audio is the original noise audio. Ground truth is the labeled mask.

model are significantly better than the three audio denoising methods in terms of SDR score. The compelling advantage of the DVAD model lies in the image segmentation section. We can maintain the crucial clean signal via a segmented mask, as shown in Fig. 6. Given a clean area in the segmented masks, the clean signal will be preserved during the ISTFT process. Therefore, converting audio denoising into an image segmentation problem can be a new stream to further improve the performance of audio denoising.

7. Conclusion

In this paper, we are the first to convert audio denoising into an image segmentation problem. We then propose a deep visual audio denoising (DVAD) network to remove the noise from a larger-scale BirdSoundsDenoising dataset. In addition, we design an audio ImageMask tool and propose to use few-shot generation to label all datasets. Extensive experimental results demonstrate that the proposed DVAD model outperforms many state-of-the-art methods. As for future work, a novel segmentation model could be developed to further improve the audio denoising performance.

Table 2: Results comparisons of different methods ($F1$, IoU, and $Dice$ scores are multiplied by 100. "–" means not applicable.

Networks	Validation				Test			
	$F1$	IoU	$Dice$	SDR	$F1$	IoU	$Dice$	SDR
U^2-Net [23]	60.8	45.2	60.6	7.85	60.2	44.8	59.9	7.70
MTU-NeT [35]	69.1	56.5	69.0	8.17	68.3	55.7	68.3	7.96
Segmenter [31]	72.6	59.6	72.5	9.24	70.8	57.7	70.7	8.52
U-Net [25]	75.7	64.3	75.7	9.44	74.4	62.9	74.4	8.92
SegNet [4]	77.5	66.9	77.5	9.55	76.1	65.3	76.2	9.43
DeepLabV3 [7]	**82.6**	**73.5**	**82.6**	**10.33**	**81.6**	**72.3**	**81.6**	**9.96**
R-CED [21]	–	–	–	2.38	–	–	–	1.93
Noise2Noise [13]	–	–	–	2.40	–	–	–	1.96
TS-U-Net [20]	–	–	–	2.48	–	–	–	1.98

8. Broad impact

The application of our proposed DVAD model is not limited to bird sound denoising. It can be easily extended to the following tasks for real-life applications.

8.1. Adaptation to speech denoising

Speech denoising is increasingly important as speech has become a predominant medium of daily communication and the main aspect of technology advancement. We use bird sounds as training samples for our model, but our model can also be applied to human speech denoising or even other non-audio signals. We applied the pre-trained DeepLabV3 model to segment the speech audio image, then converted the segmented image to get the denoised speech audio. As shown in Fig. 8, the noise in human speech audio can be significantly reduced. Therefore, our DVAD model demonstrates high-quality performance in human speech audio denoising.

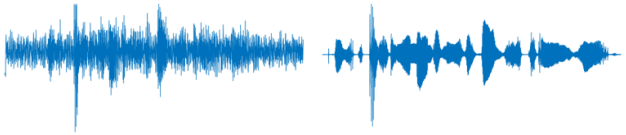


Figure 8: Example of speech denoising. Left is the original speech with noise and the right is the denoised speech audio.

8.2. Audio separation

Audio separation tasks are important in many scenarios, such as high-quality conference video production, surveillance system use, audio identification, etc. In the birds sound denoising task, we treat it as a binary image segmentation problem. For audio separation, we could treat it as a multi-class image segmentation problem. As shown in Fig. 9, we can separate two different bird sounds. We also added separated bird sounds in the supplementary material. This aspect of our model is significant when trying to detect or identify desired audio signal in a noisy mix. Hence, our DVAD model can be easily applied to the audio separation problem.

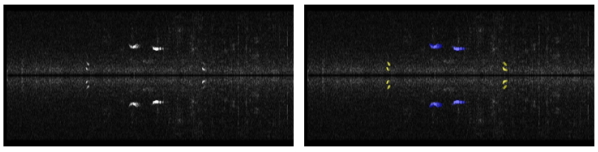


Figure 9: Example of audio separation. Left is the audio image and right is the overlay of audio image with two different segmented bird masks. Yellow color is one bird, while blue color is another bird.

8.3. Audio enhancement

Audio enhancement has always been a challenging task since the noise signal would also be enhanced if we did not properly remove noise signals. This challenge presents in many models designed for speech enhancement. In our DVAD model, we only preserve the clean signal mask. Hence, higher quality audio enhancement can be achieved. Given the denoised audio $\hat{x}_t$, we can enhance the audio by enlarging the denoised signal with $l\hat{x}_t$, where l is the number of times to enlarge the signal. As shown in Fig. 10, we can enhance the pure bird signal by $l = 200$ times. An enhanced audio example can also be found in the supplementary material. Audio enhancement has a much wider application, such as hearing aid, recording production, long-distance signal transmission like cellular communication, etc.

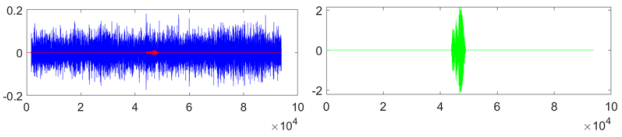


Figure 10: Example of audio enhancement. The blue line is the original noise signal, and the red line is the denoised signal. On the right, it is the enhanced signal of 200 times of signal in the red line on the left. x and y axes represent the length and magnitude of the signal, respectively.

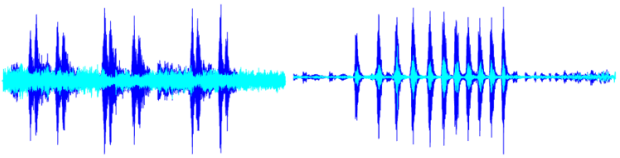


Figure 11: Two examples of noise estimation. The blue color is the noise audio, and the cyan color is the estimated noise.

8.4. Noise estimation and audio identification

In denoised audio, we could still occasionally hear the noise since the noise signal is difficult to be completely removed. Noise estimation will be useful if we can learn the patterns of noise, then we can further remove noise from the denoised audio. As shown in Fig. 11, we can estimate the noise signal by using original noise audio to subtract the clean denoised signal. Sampling these extracted noises can be used to further improve the quality of denoised signals. Noise estimation is particularly useful for model training, it could also be extended to audio identification. Noise in one scenario could become intended signals in others [38]. As each audio signal has its own pattern, learning the pattern of different audio signals can be useful to match intended signals with those in training sets. This application is important in medical applications, e.g., identifying disease-likely sound patterns from regular sound patterns.

References

[1] Houda Abouzid, Otman Chakkor, Oscar Gabriel Reyes, and Sebastian Ventura. Signal speech reconstruction and noise removal using convolutional denoising audioencoders with neural deep learning. *Analog Integrated Circuits and Signal Processing*, 100(3):501–512, 2019. 2

[2] Nasim Alamdari, Arian Azarang, and Nasser Kehtarnavaz. Improving deep speech denoising by noisy2noisy signal mapping. *Applied Acoustics*, 172:107631, 2021. 1, 2

[3] Arian Azarang and Nasser Kehtarnavaz. A review of multi-objective deep learning speech denoising methods. *Speech Communication*, 122:1–10, 2020. 2

[4] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12):2481–2495, 2017. 6, 7

[5] Steven Boll. Suppression of acoustic noise in speech using spectral subtraction. *IEEE Transactions on acoustics, speech, and signal processing*, 27(2):113–120, 1979. 1, 2

[6] Jingdong Chen, Jacob Benesty, Yiteng Huang, and Simon Doclo. New insights into the noise reduction wiener filter. *IEEE Transactions on audio, speech, and language processing*, 14(4):1218–1234, 2006. 2

[7] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018. 5, 6, 7

[8] Tian Gao, Jun Du, Li-Rong Dai, and Chin-Hui Lee. Snr-based progressive learning of deep neural network for speech enhancement. In *Interspeech*, pages 3713–3717, 2016. 3

[9] Tian Gao, Jun Du, Li-Rong Dai, and Chin-Hui Lee. Densely connected progressive learning for lstm-based speech enhancement. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5054–5058. IEEE, 2018. 3

[10] Francois G Germain, Qifeng Chen, and Vladlen Koltun. Speech denoising with deep feature losses. *arXiv preprint arXiv:1806.10522*, 2018. 2

[11] Juan S Gómez, Jakob Abeßer, and Estefanía Cano. Jazz solo instrument classification with convolutional neural networks, source separation, and transfer learning. In *ISMIR*, pages 577–584, 2018. 1

[12] John HL Hansen, Vinod Radhakrishnan, and Kathryn Hoberg Arehart. Speech enhancement based on generalized minimum mean square error estimators and masking properties of the auditory system. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(6):2049–2063, 2006. 1, 2

[13] Madhav Mahesh Kashyap, Anuj Tambwekar, Krishnamoorthy Manohara, and S Natarajan. Speech denoising without clean training data: A noise2noise approach. *Proc. Interspeech 2021*, pages 2716–2720, 2021. 6, 7

[14] Qiuqiang Kong, Haohe Liu, Xingjian Du, Li Chen, Rui Xia, and Yuxuan Wang. Speech enhancement with weakly labelled data from audioset. *arXiv preprint arXiv:2102.09971*, 2021. 1, 2

[15] Haoran Kui, Jiahua Pan, Rong Zong, Hongbo Yang, and Weilian Wang. Heart sound classification based on log mel-frequency spectral coefficients features and convolutional neural networks. *Biomedical Signal Processing and Control*, 69:102893, 2021. 1, 2

[16] Andong Li, Minmin Yuan, Chengshi Zheng, and Xiaodong Li. Speech enhancement using progressive learning-based convolutional recurrent neural network. *Applied Acoustics*, 166:107347, 2020. 1, 2, 3

[17] Bingbing Li. A principal component analysis approach to noise removal for speech denoising. In *2018 International Conference on Virtual Reality and Intelligent Systems (ICVRIS)*, pages 429–432. IEEE, 2018. 1

[18] Jae Soo Lim and Alan V Oppenheim. Enhancement and bandwidth compression of noisy speech. *Proceedings of the IEEE*, 67(12):1586–1604, 1979. 2

[19] Rainer Martin. Speech enhancement based on minimum mean-square error estimation and supergaussian priors. *IEEE transactions on speech and audio processing*, 13(5):845–856, 2005. 1, 2

[20] Eloi Moliner and Vesa Välimäki. A two-stage u-net for high-fidelity denoising of historical recordings. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 841–845. IEEE, 2022. 6, 7

[21] Se Rim Park and Jin Won Lee. A fully convolutional neural network for speech enhancement. *Proc. Interspeech 2017*, pages 1993–1997, 2017. 6, 7

[22] Mozhde Firoozi Pouyani, Mansour Vali, and Mohammad Amin Ghasemi. Lung sound signal denoising using discrete wavelet transform and artificial neural network. *Biomedical Signal Processing and Control*, 72:103329, 2022. 1, 2

[23] Xuebin Qin, Zichen Zhang, Chenyang Huang, Masood Dehghan, Osmar R Zaiane, and Martin Jagersand. U2-net: Going deeper with nested u-structure for salient object detection. *Pattern recognition*, 106:107404, 2020. 6, 7

[24] Shivangi Raj, P Prakasam, and Shubham Gupta. Multilayered convolutional neural network-based auto-codec for audio signal denoising using mel-frequency cepstral coefficients. *Neural Computing and Applications*, 33(16):10199–10209, 2021. 1, 2

[25] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 6, 7

[26] Nasir Saleem, Muhammad Irfan Khattak, Muhammad Yousaf Ali, and Muhammad Shafi. Deep neural network for supervised single-channel speech enhancement. *Archives of Acoustics*, 44, 2019. 1, 2

[27] Nasir Saleem and Muhammad Irfan Khattak. Deep neural networks for speech enhancement in complex-noisy environments. 2020. 1, 2

[28] Pascal Scalart et al. Speech enhancement based on a priori signal to noise estimation. In *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, volume 2, pages 629–632. IEEE, 1996. 1, 2

[29] Samuel Sonning, Christian Schüldt, Hakan Erdogan, and Scott Wisdom. Performance study of a convolutional time-domain audio separation network for real-time speech denoising. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 831–835. IEEE, 2020. 3

[30] Madhur Srivastava, C Lindsay Anderson, and Jack H Freed. A new wavelet denoising method for selecting decomposition levels and noise thresholds. *IEEE access*, 4:3862–3877, 2016. 1, 2

[31] Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7262–7272, 2021. 6, 7

[32] Keiichi Tamura, Akitada Omagari, and Shuichi Hashida. Novel defense method against audio adversarial example for speech-to-text transcription neural networks. In *2019 IEEE 11th International Workshop on Computational Intelligence and Applications (IWCIA)*, pages 115–120. IEEE, 2019. 1

[33] Ke Tan and DeLiang Wang. A convolutional recurrent neural network for real-time speech enhancement. In *Interspeech*, volume 2018, pages 3229–3233, 2018. 1, 2, 3

[34] Tyler Vuong, Yangyang Xia, and Richard M Stern. A modulation-domain loss for neural-network-based real-time speech enhancement. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6643–6647. IEEE, 2021. 2

[35] Hongyi Wang, Shiao Xie, Lanfen Lin, Yutaro Iwamoto, Xian-Hua Han, Yen-Wei Chen, and Ruofeng Tong. Mixed transformer u-net for medical image segmentation. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2390–2394. IEEE, 2022. 6, 7

[36] Kai Wang, Bengbeng He, and Wei-Ping Zhu. Tstnn: Two-stage transformer based neural network for speech enhancement in the time domain. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7098–7102. IEEE, 2021. 3

[37] Li Wang, Weiguang Zheng, Xiaojun Ma, and Shiming Lin. Denoising speech based on deep learning and wavelet decomposition. *Scientific Programming*, 2021, 2021. 1, 2

[38] Ruilin Xu, Rundi Wu, Yuko Ishiwaka, Carl Vondrick, and Changxi Zheng. Listening to sounds of silence for speech denoising. *Advances in Neural Information Processing Systems*, 33:9633–9648, 2020. 2, 8

[39] Yong Xu, Jun Du, Li-Rong Dai, and Chin-Hui Lee. A regression approach to speech enhancement based on deep neural networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(1):7–19, 2014. 1, 2

[40] Hongjiang Yu, Zhiheng Ouyang, Wei-Ping Zhu, Benoit Champagne, and Yunyun Ji. A deep neural network based kalman filter for time domain speech enhancement. In *2019 IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 1–5. IEEE, 2019. 3

[41] Rui-Mei Zhao and Hui-min Cui. Improved threshold denoising method based on wavelet transform. In *2015 7th International Conference on Modelling, Identification and Control (ICMIC)*, pages 1–4. IEEE, 2015. 2