IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 26 June 2023, accepted 3 July 2023, date of publication 12 July 2023, date of current version 31 August 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3294476

RESEARCH ARTICLE

DeepLabV3+ Vision Transformer for Visual Bird Sound Denoising

JUNHUI LI[1], PU WANG[1], AND YOUSHAN ZHANG[2], (Member, IEEE)
[1]Department of Mathematics, School of Science, University of Science and Technology Liaoning (USTL), Liaoning, Anshan 114051, China
[2]Department of Artificial Intelligence and Computer Science, Yeshiva University, New York City, NY 10016, USA

Corresponding author: Youshan Zhang (youshan.zhang@yu.edu)

ABSTRACT Audio denoising is a task to improve the perceptual quality of noisy audio signals. There is still residual noise after the denoising of noisy signals, which will affect the quality of audio data. Traditional and deep learning-based methods are still limited to the manual addition of artificial noise or low-frequency noise. Recently, audio denoising has been transformed into an image segmentation problem, and deep neural networks have been applied to solve this problem. However, its performance is limited to shallow image segmentation models. This paper proposes a novel vision transformer model for visual bird sound denoising, combining a pyramid transformer and DeepLabV3+ network (named PtDeepLab) to filter out the noise. The proposed PtDeepLab model is based on the pyramid transformer, which generates long-range and multi-scale representations. The PtDeepLab model can achieve intuitive noise reduction in audio, which helps to separate clean audio from the mixture signal. Extensive experimental results showed that the proposed model has a better denoising performance than state-of-the-art methods.

INDEX TERMS Audio denoising, transformer, DeepLabV3+.

I. INTRODUCTION

Audio denoising is a long-standing challenge for many tasks (e.g., teleconferences, the speech-to-text function in social media, and hearing aid) [1]. With the popularity of the Internet in recent years, audio signals are widely used in our life for information transmission. In the process of information transmission, all kinds of noise will affect the clarity of the audio. The maintenance of speech signal transmission quality and retaining as much useful information as possible are the main purposes of audio denoising. Over the last decade, audio denoising research has shown that a viable solution is to build a noise estimation generative model and use it to recover intelligible audio signals with better quality from noisy audio signals [2], [3], [4]. However, these methods with added artificial noise or lower denoising quality have their limitations and may not be efficient for speech processing.

Audio denoising can significantly improve audio quality. Typically, traditional statistical methods and deep learning methods are used to reduce noise and separate audio. There are six different kinds of methods, including both traditional and deep learning models: (1) optimal FIR filter, (2) spectrum subtraction, (3) minimum mean square error, short-time spectral amplitude estimator (MMSE-STSA) [36], (4) wavelet noise reduction based on image noise reduction, (5) processing-based image noise reduction, (6) noise reduction based on deep learning [13]. In recent years, some researchers have shown that image processing-based noise reduction methods with deep learning models outperform traditional methods. Deep learning-based audio-denoising algorithms have attracted wide attention and revolutionized the domain of audio denoising. By learning a deep nonlinear network structure, deep neural networks (DNNs) have superior potential for complicated nonlinear mapping problems and can be used for audio denoising. Different deep learning-based audio-denoising approaches can often be categorized into two main groups: the spectral mapping approach and the mask mapping approach. By ignoring the structural features of the speech spectrum and the long contextual relationships between adjacent frames, these methods often lead to spectral artifacts and speech distortion in high-frequency bands. Bird sounds play an essential role in animal sound

The associate editor coordinating the review of this manuscript and approving it for publication was Hui Ma.

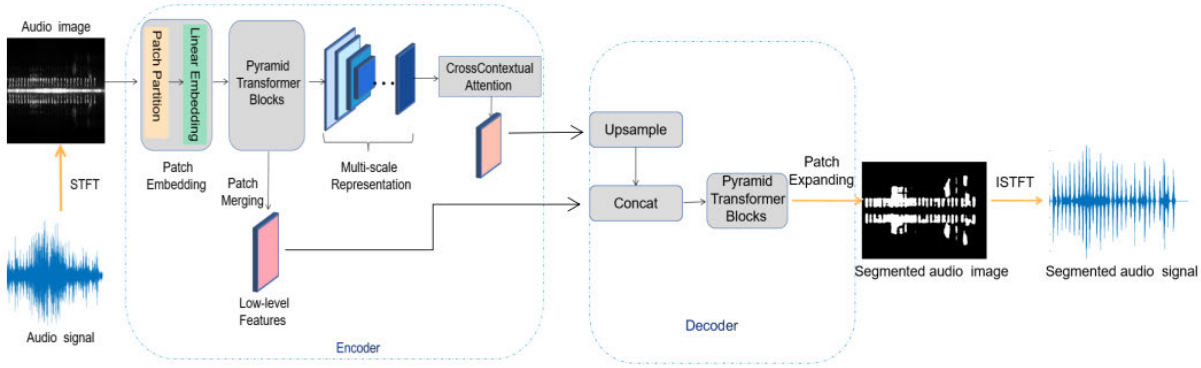

FIGURE 1. The overall progress of our proposed vision transformer framework. Different modules are marked with different color blocks. The architecture of the main body extends the encoder-decoder structure of the DeepLabV3+ box based on a pyramid transformer.

recognition. Animal sound classification usually has three processing steps [56]: signal preprocessing, feature extraction, and classification. Signal preprocessing mainly includes signal segmentation and denoising. However, many audio signals are directly collected in nature and obtained in a relatively noisy environment, making it necessary to artificially reduce the noise from these biological recordings that are directly obtained from nature. If we cannot accurately filter out the noise signals, it will lead to sound distortion and affect the recognition results of animal sounds [5]. This paper mainly investigates how to reduce noise components in bird sounds. Influenced by the neural network [6], [7], [8], the audio signal is converted into images via the Short-Time Fourier Transform (STFT) method. The image is then segmented to achieve the purpose of noise reduction via the deep learning method [9], [10], using samples from the natural environment to perform audio noise reduction.

In this paper, we aim to propose a robust segmentation model that is able to separate the clean audio from the mixture audio. As shown in Fig. 1, our proposed model consists of three key modules: the STFT module, a PtDeepLab model, and the Inverse Short-Time Fourier Transform (ISTFT) module. We convert the audio signal to an audio image using the STFT module and then train the PtDeepLab model to segment the clean audio signal region. After obtaining the clean audio signal region, we finally applied the ISTFT module to reconstruct the denoised audio.

Our contributions are three-fold:

- We develop a novel deep visual transformer in the visual bird audio denoising model that transfers the audio denoising to an image segmentation problem to achieve the purpose of audio denoising by removing the noise region in the audio image.
- We propose a transformer-based encoder-decoder architecture to capture and fuse the multi-scale representation. The proposed PtDeepLab is based on a pyramid transformer and DeepLabV3+ box, which is a transformer-based framework that achieves audio image segmentation with varying levels of resolution.
- We achieved new state-of-the-art audio denoising results on the BirdSoundDenoising dataset, demonstrating the effectiveness of the proposed method and the enhanced ability to learn data and features.

II. RELATED WORK

There are various audio denoising methods, which can be divided into three categories according to the signal representation type: time-domain, frequency-domain, and time-frequency domain. The vanilla audio denoising processes are as follows: extract the relevant features and convert them into the correct format, then define the Fourier window to calculate the Fourier transform of the signal, train the network to produce an estimated value, and optimize the average variance between the output and the target signal. Our work is related to three major research directions, and we highlight some representative methods that are closely related to our work.

A. TRADITIONAL AUDIO DENOISING METHODS

They mainly rely on estimating audio statistics [22]. Statistical methods like the Gaussian mixture model can be used to build a denoising model of interest and recover clean audio from the noisy input signal. Many methods use the STFT and the ISTFT [17], which are time-domain algorithms, to solve the audio enhancement problem and treat audio enhancement as a filtering problem. The denoising performance can be improved by the Wiener filter [11] or the LSA estimator [12]. Gradolewski et al. [58] used Gaussian white noise for simulation, using a wavelet-denoised algorithm to filter the real phonocardiography (PCG) signal interference generated from signals recorded by a mobile device in a noisy environment. Linden et al. [25] decomposed a spectral graph into two matrices: the spectral basis matrix and the encoding matrix. Spectral bases belonging to the same source are then grouped according to the periodicity of the encoded information. Finally, the different sound sources are reconstructed based on the clustering of the basis matrix and the corresponding encoding information. And then, the noise components are

removed to facilitate more accurate monitoring of biological sounds. Zha et al. [26] proposed a novel rank residual constraint (RRC) model for the rank minimization problem and applied it to image restoration tasks. In addition, some new image restoration approaches were also presented, such as the simultaneous nonlocal self-similarity priors method [27] and the group sparsity residual constraint with non-local priors method [28]. Haider et al. [29] proposed a filter-based denoising scheme using the signal-to-noise ratio, which is mainly used to simulate different levels of addictive Gaussian noise. For example, the frequency range is usually between 1 kHz and 12 kHz [30], [31]. Some traditional methods of sound denoising have difficulty tracking target sound sources in multiple sources, which means that they cannot handle long-term contexts [14], [34].

B. DEEP LEARNING METHODS

For the comparison of the deep learning model and the traditional supervised methods, Alamdar et al. [15] applied a full convolutional neural network (FCN) to denoise the audio with only noise samples, reflecting the superiority of the deep learning model. Germain et al. [16] trained the FCN using deep feature loss, and their model can interfere with background noise to suppress the noisy signals. Xu et al. [47] introduced a deep learning model with automatic speech denoising, which can better capture noise patterns. Saleem et al. [55] used an ideal binary mask (IBM) and the training DNNs to estimate the IBM, which is important for the audio enhancement of complex noise. The result also showed that DNNs have a better ability to learn data and features from a few samples. Xu et al. [35] proposed a DNN-based supervised method to enhance the audio by finding a mapping function between noisy and clean audio samples. Madhav [49] proposed a Noise2Noise approach to tackle the problem of the heavy dependence on clean speech data. Takuya et al. [45] proposed a training strategy that does not require clean signals. Moreover, Tao et al. [42] presented a method called Neighbor2Neighbor to train an effective image-denoising model without only noisy images. Aswin et al. [43] proposed self-supervised learning methods as a solution to both zero- and few-shot personalization tasks. Sonining et al. [37] investigated the performance of such a time-domain network (Conv-TasNet) for speech denoising in a real-time setting, comparing various parameter settings.

C. VISION TRANSFORMER

Transformer was originally proposed for natural language processing (NLP) tasks. Transformers, which are currently state-of-the-art across domains, including natural language processing and computer vision (CV), have gone viral in the field of speech processing [44]. Transformer has also been adapted for audio processing with CNNs. Some authors stack a transformer on top of a CNN, and they combine a transformer and a CNN in each model block. Other efforts combine CNNs with simpler attention modules. Kong et al. [19] presented CleanUNet, which is based on an encoder-decoder architecture combined with self-attention blocks. It has to be mentioned that Vision Transformer (ViT) is the first transformer-based approach that can match or even surpass CNNs in image classification. Many variants of visual transformers have also been proposed recently. Liu et al. [53] proposed a new vision transformer, called Swin Transformer, whose architecture has the flexibility to model at various scales and has linear computational complexity concerning image size. Chen et al. [20] proposed CrossVit, a dual-branch transformer to combine image patches (i.e., tokens in a transformer) of different sizes to produce stronger image features. Gu et al. [24] proposed HRViT, which enhances ViTs' ability to learn semantically rich and spatially precise multi-scale representations by integrating high-resolution multi-branch architectures with ViTs. Recently, some transformer methods have been used to solve audio processing problems. Gong et al. [21] built the Audio Spectrogram Transformer (AST), the first convolution-free, a purely attention-based model for audio classification.

In this work, we focus on a deep learning-based audio denoising method using nature datasets. Our main inspiration for this work is based on Zhang and Li [18], who were the first to convert audio denoising into a visual image segmentation problem. They first converted bird audio to images using the STFT and proposed to segment the clean audio areas and remove the noisy areas. Finally, they applied ISTFT to convert segmented, clean audio images into audio to realize the purpose of audio denoising. However, they did not propose any new segmentation models and left the space to further improve the performance of their proposed BirdSoundDenoising datasets. Priyadarshani et al. [57] have described a combination of denoising methods using wavelet packet decomposition and band-pass or low-pass filtering. Their presented experiments demonstrate an order of magnitude improvement over the noise reduction recorded by natural bird noise. However, their model still has lower performance in large-scale bird noise datasets. In this paper, we can consider deleting the noise from the frequency domain. If we can remove the noise areas from the frequency domain, we can achieve the purpose of noise reduction. Inevitably, there are always specific noise frequencies interspersed with the frequency of bird sounds that cannot be removed. Therefore, the denoising method of removing noise from the frequency domain has certain limitations. In view of this problem, we propose a DeepLabV3+ Vision Transformer method to reduce noise areas and transform the audio noise reduction problem into an image segmentation problem. After receiving a noisy input signal, we will convert it into an image and remove the noisy areas to extract a clean bird sound signal.

III. METHODS

A. PROBLEM

A noisy audio signal $y(t)$ can be typically expressed as:

$$y(t) = x(t) + \varepsilon(t) \tag{1}$$

where $x(t)$ and $\varepsilon(t)$ denote clean audio and additive noise signals of time index t, respectively [14]. A sequence of noisy signal and clean signal are defined as $Y = \{y_i\}_{i=1}^{N}$ and $X = \{x_i\}_{i=1}^{N}$, where N is the total number of audios. The goal of audio denoising is to extract the clean audio component X from the mixture audio signal Y by learning a mapping $\mathcal{F}$ and minimize the approximation error between the estimated denoised audio $\mathcal{F}(Y)$ and clean audio X. In our paper, we also work on converting the audio denoising to an image segmentation task. Given the audio images $\mathcal{I} = \{I_i\}_{i=1}^{N}$ corresponding to Y, we aim to minimize the error between the predictions of our segmentation model PtDeepLab ($\mathcal{I}$) and its ground truth labeled masks $U = \{u_i\}_{i=1}^{N}$.

B. MOTIVATION

Most traditional filtering methods are limited to window-adding or masking operations in the frequency domain or time domain. Because of the strong time-frequency coupling between the audio signal and noise, these filtering methods are difficult to use to achieve effective signal and noise separation. Many existing deep audio denoising methods use clean audio signals as output signals or study the magnitude spectrum of the image to denoise. However, these methods can be constrained by computing power or limited filtering image areas, leading to low denoising performance. The scientific goal of this paper is to develop a novel, fully automatic deep-learning denoising model that can discover differences between noisy and clean signal regions by digging into audio images. If the clean signal area can be successfully segmented, the goal of audio denoising can be achieved. In summary, given a noisy input signal, we aim to build a deep learning model that can extract clean signals and return them to the user.

C. PRELIMINARY

In our model, we aim to obtain the raw images for each bird sound by the STFT and reconstruct the denoised bird's sound based on the segmented bird sound image by the ISTFT.

1) STFT THEORY

The Short-Time Fourier Transform (STFT) and Inverse Short-Time Fourier Transform (ISTFT) are widely used in speech analysis and processing. They are suitable for slow signal and time-varying signal spectrum analysis [22]. The audio signal is non-stationary in most cases, meaning that the mean and variance of the signal are not constant over time. Therefore, it does not make much sense to calculate the Fourier transform on the whole audio signal, so the Fourier transform with window length and jump size values is proposed [23]. In this method, the audio signal is first divided into frames. Then, each frame of the audio signal can be intercepted from various fixed signal waveforms by the Fourier transform, and the short-term spectrum of each frame is an approximation of the spectrum value of the smooth signal waveform.

2) AUDIO IMAGE CONSTRUCTION

STFT is a function of time t and frequency f, which shows how the frequency of the speech signal changes with time. Fig. 1 shows the conversion from the bird sound audio to its audio image. We can obtain the audio image (I) after implementing the STFT calculation and the following equation,

$$I = abs(STFT_y(t, f)) \tag{2}$$

where $STFT_y(t, f)$ is the coefficient of STFT and abs takes the absolute value from the complex frequency domain $\mathcal{O}$.

3) DENOISED AUDIO RECONSTRUCTION

After we remove the noise areas from the frequency domain, we can apply ISTFT to reconstruct the denoised audio signal. Firstly, we need to filter out noise areas in $\mathcal{O}$ from the segmentation model. The new frequent domain $\mathcal{O}'$ is as follow Eq. (3). More details of the ISTFT process are shown in the Fig. 1.

$$\mathcal{O}' = \mathcal{O}, \quad and \quad \mathcal{O}'[\hat{\tilde{u}} < 1] = 0 \tag{3}$$

where $\tilde{u}$ = PtDeepLab (I) is the predicted mask of the segmentation model given the input image I and we finally reconstruct the denoised audio as follows:

$$\hat{y}(t) = ISTFT(\mathcal{O}') \tag{4}$$

D. PROPOSED MODEL: PTDEEPLAB

The transformer method has gained wide attention in natural language processing and computer vision in recent years thanks to global information modeling derived from the self-attention mechanism [24]. Previous studies have demonstrated that both local and global features are essential for depth models in dense prediction, such as the segmentation of complex structures. In this section, we will introduce how to directly apply a transformer to image patch feature representation coding and elaborate on the overall framework of PtDeepLab.

1) ENCODER-DECODER ARCHITECTURE

An overview of the PtDeepLab model is depicted in Fig. 1. Our proposed PtDeepLab extends DeepLabV3+ [54], [59] by employing pyramid transformer blocks in the encoder and decoder. Specifically, the encoder module encodes the input image into a highly representative space by applying a pyramid transformer on multiple scales to encode multi-scale contextual information. At the same time, the simple yet effective decoder is utilized as a series of pyramid transformer blocks with path-expanding operations applied to reach the full resolution of images. Given an audio image $I \in \mathbb{R}^{H \times W \times C}$ with a spatial resolution of $H \times W$ and C number of channels, our goal is to predict the corresponding mask given the input image I. After segmenting the clean sound areas in the audio image, we could remove the noise from the audio signal.

In our implementation, given an input audio image of size $H \times W \times 1$, we first divide it into non-overlapping patches of

size 8×8, and thus the feature dimension of each patch is $8 \times 8 \times 1 = 64$. Each patch is treated as a ''token'', whose feature is set as a concatenation of the raw pixel values. The pyramid transformer is applied to encode both local semantic and long-range contextual representations. To construct Pt Spatial Pyramid Pooling (PSPP), the pyramid transformer block is designed to capture multi-scale information representation. The obtained multi-scale contextual representations are fused into the decoding module through a cross-contextual attention mechanism. The cross-contextual attention block consists of a channel attention operation and a spatial attention operation, and they are applied to the tokens (derived from each level of the pyramid) to formulate the multi-scale interaction [32]. In the final decoding process, we first upsample bilinearly the extracted encoder high-level features and concatenate them with the low-level features from the pyramid transformer backbone in the encoder to update the feature representation.

2) PATCH EMBEDDING

An image $I \in \mathbb{R}^{H\times W\times C}$ is split into a sequence of patches $\zeta = [\zeta_1, \ldots, \zeta_N] \in \mathbb{R}^{N\times P^2\times C}$, where (P, P) is the patch size, $N = HW/P^2$ is the number of patches. We map the vectorized patches ζ_p into a latent D-dimensional embedding space using a trainable linear projection to produce a sequence of patch embeddings. To encode the patch spatial information, we learn specific position embeddings $Pos = [Pos_1, \ldots, Pos_N] \in \mathbb{R}^{N\times D}$, which are added to the patch embeddings to retain positional information as follows:

$$\mathcal{Z} = [\zeta_1 E; \zeta_2 E, \cdots; \zeta_N E] + E_{Pos} \tag{5}$$

where $E \in \mathbb{R}^{(P^2\cdot C)\times D}$ is the patch embedding projection, and $E_{Pos} \in \mathbb{R}^{N\times D}$ denotes the position embedding.

3) PYRAMID TRANSFORMER BLOCK

Because of the traditional transformers' single-scale, low-resolution representations, it is difficult for ViTs to implement dense prediction tasks such as semantic segmentation and effectively leverage the rich transformer layers in the encoder for excavating helpful multi-modal context. In addition, these methods incur high computational and memory costs due to the global self-focus mechanism. Pyramid transformer is designed to alleviate this problem [51], [52]. The key design feature of the pyramid transformer is to design a progressive shrinking pyramid and spatial-reduction attention (SRA). SRA is a substitute for a multi-head self-attention (MSA) module in the transformer block. Thus, each pyramid transformer block comprises an attention layer and a feed-forward layer with a LayerNorm (LN) layer, a two-layer MLP, and GELU nonlinearity. The spatial-reduction attention (SRA) module is applied in series in the transformer block, as depicted in Fig. 2. With such a spatial-reduction attention module, the successive pyramid transformer blocks can be

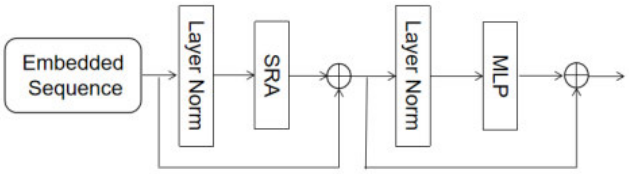

FIGURE 2. Schematic of the Transformer layer used in this work.

expressed as:

$$SRA(Q, K, V) = Concat(head_0, \cdots, head_{N_i})W^O \tag{6}$$

$$head_j = Attention(QW_j^Q, SR(K)W_j^K, SR(V)W_j^V) \tag{7}$$

where $Concat(\cdot)$ is the concatenation operation. W_j^Q, W_j^K, $W_j^V \in \mathbb{R}^{C_i\times d_{head}}$ and $W^O \in \mathbb{R}^{C_i\times C_i}$ are linear projection parameters. N_i is the head number of the attention layer in the stage i. $SR(\cdot)$ is the operation for reducing the spatial dimension of the patch embedding, which is written as:

$$SR(\mathcal{Z}) = Norm(Reshape(\mathcal{Z}, R_i)W^S) \tag{8}$$

Here, $\mathcal{Z} \in \mathbb{R}^{(H_iW_i)\times C_i}$ represents a patch embedding, and R_i denotes the reduction ratio of the attention layers in stage i. $W^S \in \mathbb{R}^{(R_i^2C_i)\times C_i}$ is a linear projection that reduces the dimension of the patch embedding to C_i. Norm$(\cdot)$ refers to layer normalization.

With such a spatial-reduction scheme, consecutive pyramid transformer blocks can be formulated as:

$$\begin{aligned}\hat{\mathcal{Z}}^l &= SRA(LN(\mathcal{Z}^{l-1})) + \mathcal{Z}^{l-1}\\ \mathcal{Z}^l &= MLP(LN(\hat{\mathcal{Z}}^l)) + \hat{\mathcal{Z}}^l\end{aligned} \tag{9}$$

where $\hat{\mathcal{Z}}^l$ and $\mathcal{Z}^l$ denote the output features of the SRA module and the MLP module for block L_i, respectively. The self attention as in the ViTs Transformer [33] is computed according to:

$$Attention(\mathcal{Q}, \mathcal{K}, \mathcal{V}) = SoftMax(\frac{\mathcal{Q}\mathcal{K}^T}{\sqrt{d}})\mathcal{V} \tag{10}$$

where $\mathcal{Q}, \mathcal{K}, \mathcal{V} \in R^{N\times d}$ are the query, key and value matrices; d is the query/key dimension.

4) ENCODER

Different from the CNN backbone networks, which use different convolution steps to obtain multi-scale feature maps, our model uses stacked pyramid transformer modules as the encoder. Furthermore, the pyramid transformer encoder has four stages, which comprise some blocks at each stage. After that, the embedded patches, along with a position embedding, are passed into some successive pyramid transformer blocks with L_i layers to generate hierarchical representations. Stage 1, stage 2, stage 3, and stage 4 have layers of 3, 4, 6, and 3, respectively. In the beginning, our PtDeepLab encoder first divides an input image into $\frac{H}{8} \times \frac{W}{8}$ patches and feeds the flattened patches to a linear projection. The output is reshaped

to a feature map M_1 of size $\frac{H}{8} \times \frac{W}{8} \times C_1$. To maintain the hierarchical structure of the encoder, a patch merging layer is utilized to decrease the resolution of feature representations by a factor of 2 at the end of each stage. In the same way as stage 1, using the feature map from the previous stage as input, we obtain the following feature maps: M_2, M_3, and M_4, whose strides are 16, 32, and 64 pixels with respect to the input image. Then, through stacking a series of pyramid transformer blocks, the spatial dimension of the feature graph is gradually reduced (similar to the CNN encoder), and the feature dimension is increased. The results are then fed into the PSPP module to capture multiscale representations.

5) PT SPATIAL PYRAMID POOLING

The PSPP block showed that regions of an arbitrary scale could be accurately and efficiently classified by resampling convolutional features extracted at a single scale. We have implemented a variant that uses multiple parallel atrous convolutional layers with different sampling rates. The extracted features for each sampling rate are further processed in separate branches and fused to generate the final result. To capture contextual information at multiple scales, compensate for spatial representations, and produce multiscale representations, our PtDeepLab model utilizes a PSPP module that replaces the pooling operation with atrous convolution. With stacked pyramid transformer blocks and subsequent patch merging layers (similar to the continuous downsampling operation in the CNN encoder), the spatial resolution of the deep features extracted by the pyramid transformer block is greatly reduced. Specifically, PtDeepLab applies several parallel convolution operations with multiple different rates, making it possible to construct a feature pyramid. To model a pooling operation such as a pure transformer, we create a Pt spatial pyramid pooling (PSPP) block with different window sizes to capture the multi-scale representation. In our design, local and global information are captured by the smaller window and the larger window, respectively. The final multi-scale representations are then fed to the cross-context attention module, where the common representations are fused and captured using nonlinear techniques.

6) CROSS-CONTEXTUAL ATTENTION

In our model, a cross-attention module is applied to model multi-scale interactions and incorporate pyramid features. We assume that each level of the pyramid represents the object of interest on different scales, thus concatenating all these features into a new dimension. We adopt a multi-scale representation $z_{all}^{P\times NC} = [z_1||z_2\cdots||z_M]$, where $||$ shows the concatenation operation, N and C indicate the number of tokens and embedding dimension, and M denotes the level number of the pyramid. Second, considering the global representation of each channel and the channel representation between pyramid levels, a proportional attention module is applied to adaptively emphasize the contribution of each feature map and outperform the less disparate features. A channel attention operation ATT_c and a spatial attention operation ATT_s can be formulated as:

$$\begin{aligned} \mathcal{Z}' &= ATT_c(\mathcal{Z}) \otimes \mathcal{Z} \\ \mathcal{Z}'' &= ATT_s(\mathcal{Z}') \otimes \mathcal{Z}' \end{aligned} \tag{11}$$

The channel attention operation ATT_c can be written in the following equation.

$$ATT_c(z) = \sigma(\mathcal{W}_1(P_{max}(z)) + \mathcal{W}_2(P_{avg}(z))) \otimes z \tag{12}$$

where z is the input feature map and σ is the Sigmoid activations, P_{max} and P_{avg} denote adaptive maximum pooling and adaptive average pooling functions, respectively. $\mathcal{W}_i$, $i \in 1, 2$ shares parameters and consists of a convolutional layer with 1×1 kernel size to reduce the channel dimension 12 times, followed by a ReLU layer and another 1×1 convolutional layer to recover the original channel dimension.
The spatial attention operation ATT_s can be formulated as:

$$ATT_s(z) = \sigma(\mathcal{K}(Concat(R_{max}(z), R_{avg}(z)))) \otimes z \tag{13}$$

where R_{max} and R_{avg} represent the maximum and average values obtained along the channel dimension, respectively. $\mathcal{K}$ is a 1×1 convolutional layer with padding set to 0.

7) DECODER

In the decoder process, the obtained features corresponding to the attention module are first up-sampled by the pyramid transformer block with a factor of 4 and then concatenated with the low-level features. The scheme of concatenation of shallow and deep features together reduces the loss of spatial detail with the help of the subsampling layer. Finally, a series of cascaded pyramid transformer blocks with path extension operations are applied to achieve the full resolution of $H \times W$.

8) OBJECTIVE FUNCTION

We use the Dice loss function to optimize our proposed PtDeepLab model.

$$Diceloss = 1 - 2 \times \frac{u \cap \widetilde{u}}{u + \widetilde{u}} \tag{14}$$

The overall training algorithm is shown in Alg. 1.

IV. DATASETS

A. DATASETS AND COMPARED MODELS

To test our proposed model, we show its performance on the BirdSoundsDenoising dataset.

1) BIRDSOUNDSDENOISING

The BirdSoundsDenoising dataset, which has 14,120 audio signals from one second to fifteen seconds and contains 10,000/1,400/2,720 in training, validation, and testing, respectively, is a large-scale dataset of bird sounds. [18]. Unlike many audio-denoising datasets, which have manually added artificial noise, this dataset contains many natural noises, including wind, waterfalls, rain, etc.

Algorithm 1 DeepLabV3+ Vision Transformer for Visual Bird Sound Denoising. Batch of Audio Images: $B(\mathcal{I}) = \{\mathcal{I}^1, \ldots, \mathcal{I}^{n_B}\}$, and Their Labeled Mask Images $B(M) = \{M^1, \ldots, M^{n_B}\}$, Where n_B Is the Total Number of Batch. II Is the Number of Iterations and k Is One Batch

Input: Noise audio signals $Y = \{y_i\}_{i=1}^N$ and labeled mask images $M = \{m_i\}_{i=1}^N$, where N is the total number of audios.
Output: Denoised audio signals $\mathcal{F}(Y)$
Generate audio images $\mathcal{I}$ using Eq. (2)
for $iter = 1$ **to** II **do**
  for $k = 1$ **to** n_B **do**
    Derive batch-wise data: $\mathcal{I}^k$ and M^k sampled from $\mathcal{I}$ and M
    Optimize our segmentation model PtDeepLab using Eq. (14)
  end for
end for
Get the clean frequency domain using Eq. (3)
Output the denoised audio signals using Eq. (4)

2) BASELINE MODELS

We compare our results with other nine models, including $U^2 - Net$ [46], MTU-Net [41], Segmenter [39], U-Net [40], SegNet [38], DVAD [18], R-CED [48], Noise2Noise [49] and TS-U-Net [50]. For a fair comparison, we evaluate these models for both validation and test datasets.

B. IMPLEMENTATION DETAILS

We implement our model using the PyTorch framework with an RTX A6000 GPU to speed up the computation. It took less than 0.5 seconds per audio image during the inference, while it took around 2 weeks of GPU time to train our model. The hyperparameter details are as follows: We used the AdamW optimizer for pyramid transformer networks to update the network parameters. The learning rate is set to 1e-4, and the weight decay is adjusted to 1e-4 too. Further, we resize the input images to $512 \times 512 \times 3$ with a mini-batch size of 8 for 100 epochs. Nine different state-of-the-art segmentation methods use the same training settings as the above PtDeepLab.

V. RESULTS

In this section, we compare our PtDeepLab model with existing methods in terms of learning ability, capability, and qualitative results.

A. EVALUATION METRICS

We employ three widely-used evaluation metrics, including F1, IoU, and Dice to evaluate the performance of image segmentation [18]. For audio denoising, we use signal-to-distortion ratio (SDR) to evaluate our model using Eq. (15). The higher these four metrics, the better of segmentation model is.

$$SDR = 10log_{10}\frac{||u||^2}{||\widetilde{u} - u||^2} \tag{15}$$

TABLE 1. Results comparisons of different methods (*F*1, *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 [46]	60.8	45.2	60.6	7.85	60.2	44.8	59.9	7.70
MTU-NeT [41]	69.1	56.5	69.0	8.17	68.3	55.7	68.3	7.96
Segmenter [39]	72.6	59.6	72.5	9.24	70.8	57.7	70.7	8.52
U-Net [40]	75.7	64.3	75.7	9.44	74.4	62.9	74.4	8.92
SegNet [38]	77.5	66.9	77.5	9.55	76.1	65.3	76.2	9.43
DVAD [18]	82.6	73.5	82.6	10.33	81.6	72.3	81.6	9.96
PtDeepLab	**83.4**	**75.9**	**83.4**	**10.49**	**83.1**	**75.4**	**83.0**	**10.43**
R-CED [48]	–	–	–	2.38	–	–	–	1.93
Noise2Noise [49]	–	–	–	2.40	–	–	–	1.96
TS-U-Net [50]	–	–	–	2.48	–	–	–	1.98

TABLE 2. Ablation results.

Model	$F1$	IoU	$Dice$	SDR
DeepLabv3	82.6	73.5	82.6	10.33
Pyramid-transformer	79.4	72.4	80.5	10.14
Full Model	**83.4**	**75.9**	**83.4**	**10.49**

B. EXPERIMENTS

In our experiments, we compare our model with another nine different baseline models on the BirdSoundsDenoising dataset to evaluate the performance of our proposed model. The first six selected segmentation models in Tab. 1 also have an encoder-decoder structure. The encoder-decoder structure of these comparison methods has a similar architecture to ours. To demonstrate the superiority of our proposed model, we also compare it with three other audio-denoising methods.

C. PERFORMANCE COMPARISONS

Fig. 3 shows the comparisons of six different segmentation models. The segmentation mask of the PtDeepLab model has better performance than that of other models. The compared results are tabulated in Tab. 1, which compares our best results with those of previous state-of-the-art models. Four commonly used objective performance metrics from section V-A are considered to evaluate the effectiveness of the developed PtDeepLab method. We can observe that our model outperforms other competitors in terms of evaluation metrics (F1, IoU, and Dice scores) for the BirdSoundsDenoising dataset among all segmentation models. It is worth noting that three of the audio denoising methods (R-CED, Noise2Noise, and TS-U-Net) performed relatively lower than all other segmentation models. Furthermore, the SDR score of our PtDeepLab model achieves the highest value between the average SDRs of all bird sounds in the validation and test datasets. The comparisons of raw bird audio, ground truth labeled denoised audio, and denoised audio from other models are shown in Fig. 4. Our model is also closer to the

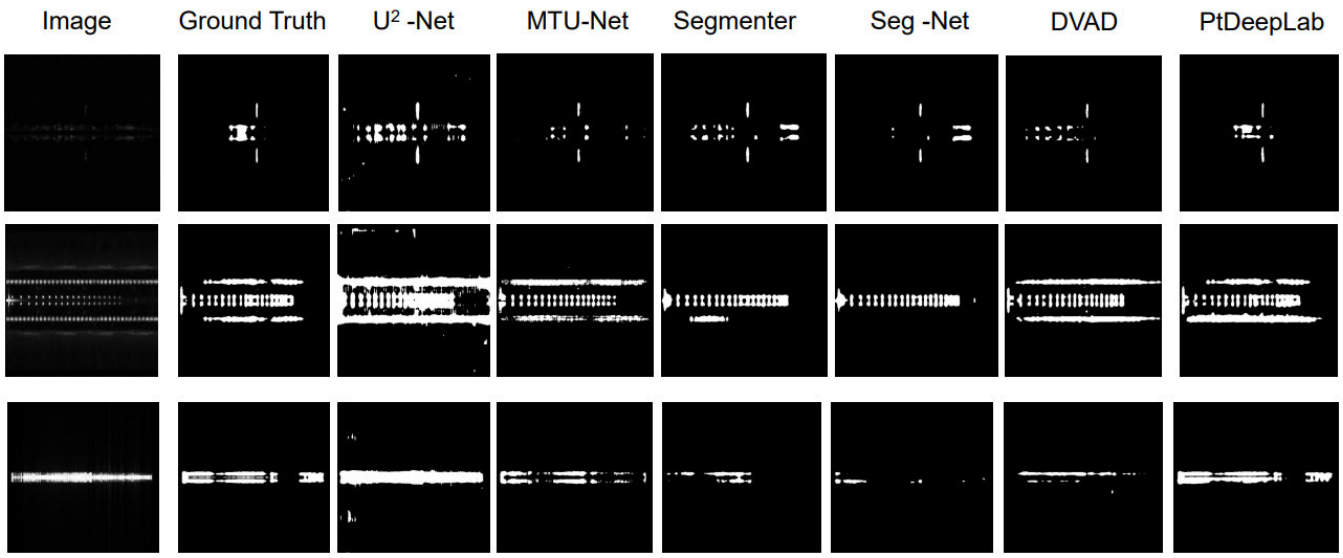

FIGURE 3. Segmentation results comparisons. Leftmost column is the original audio image. Ground truth is the labeled mask.

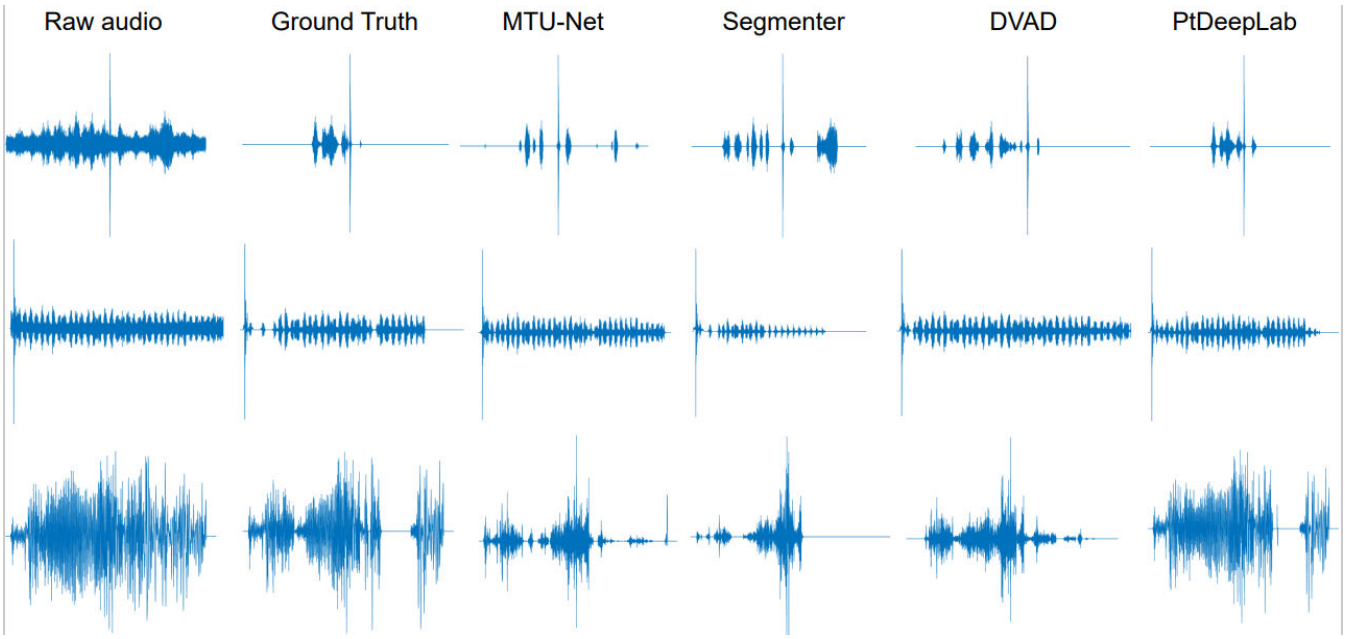

FIGURE 4. Denoising results comparisons. Raw audio is the original noise audio.

labeled, denoised signal. As a result, these benchmarks validate the effectiveness of our approach in segmenting images, and our model improves the audio-denoising performance of BirdSoundDenoising datasets.

VI. DISCUSSION

Compared to the CNN-based DeepLab model, our approach produces better segmentation results and improves the representation ability of the model in context patterning by probing features at multiple scales to attain multi-scale information. In addition, among six different state-of-the-art segmentation models and three deep audio denoising methods, one obvious advantage of our model is its higher performance than other methods. To further validate the effectiveness of the proposed method, we performed an ablation analysis. In Tab. 2, we can observe that ablation results for a DeepLabv3 model and a pyramid transformer model are compared to the full PtDeepLab model, which demonstrates the effectiveness of the improved transformer. The compelling advantage of our model lies in the image segmentation section. We can maintain the crucial clean signal via a segmented mask, as shown in Fig 3. Therefore, the novel proposed segmentation model

successfully improves the performance of the already proposed birdsoundnoise dataset.

VII. CONCLUSION

In this paper, we propose a novel DeepLabV3+ Vision Transformer model to remove the noise from the large-scale BirdSoundsDenoising dataset. Based on the DeepLab framework, the main body of PtDeepLab utilizes the pyramid transformer backbone as an encoder to explicitly extract more powerful and robust features. Extensive experimental results demonstrate that the proposed model outperforms many state-of-the-art methods, including CNN-based self-attention methods. As for future work, we look forward to evaluating DeepLavbV3+ Vision Transformer on other dense prediction vision tasks to stimulate more novel ideas for solving the visual task and demonstrate the strength of this model as a vision backbone.

REFERENCES

[1] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-33, no. 2, pp. 443–445, Apr. 1985.

[2] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-27, no. 2, pp. 113–120, Apr. 1979.

[3] P. Scalart and J. V. Filho, "Speech enhancement based on *a priori* signal to noise estimation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. Conf.*, May 1996, pp. 629–632.

[4] Y. Ephraim and H. L. Van Trees, "A signal subspace approach for speech enhancement," *IEEE Trans. Speech Audio Process.*, vol. 3, no. 4, pp. 251–266, Jul. 1995.

[5] N. Krishnamurthy and J. H. L. Hansen, "Babble noise: Modeling, analysis, and applications," *IEEE Trans. Audio, Speech, Language Process.*, vol. 17, no. 7, pp. 1394–1407, Sep. 2009.

[6] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, "PANNs: Large-scale pretrained audio neural networks for audio pattern recognition," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 28, pp. 2880–2894, 2020.

[7] J. Hou, S. Wang, Y. Lai, Y. Tsao, H. Chang, and H. Wang, "Audio-visual speech enhancement using multimodal deep convolutional neural networks," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 2, no. 2, pp. 117–128, Apr. 2018.

[8] A. A. Nugraha, A. Liutkus, and E. Vincent, "Multichannel audio source separation with deep neural networks," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 24, no. 9, pp. 1652–1664, Sep. 2016.

[9] I. R. I. Haque and J. Neubert, "Deep learning approaches to biomedical image segmentation," *Informat. Med. Unlocked*, vol. 18, Jan. 2020, Art. no. 100297.

[10] O. Habimana, Y. Li, R. Li, X. Gu, and G. Yu, "Sentiment analysis using deep learning approaches: An overview," *Sci. China Inf. Sci.*, vol. 63, no. 1, pp. 1–36, Jan. 2020.

[11] H. Lin, Y. Song, H. Wang, L. Xie, D. Li, and G. Yang, "Multimodal brain image fusion based on improved rolling guidance filter and Wiener filter," *Comput. Math. Methods Med.*, vol. 2022, pp. 1–17, Oct. 2022.

[12] I. Cohen, "Optimal speech enhancement under signal presence uncertainty using log-spectral amplitude estimator," *IEEE Signal Process. Lett.*, vol. 9, no. 4, pp. 113–116, Apr. 2002.

[13] J. Xie, J. G. Colonna, and J. Zhang, "Bioacoustic signal denoising: A review," *Artif. Intell. Rev.*, vol. 54, no. 5, pp. 3575–3597, Jun. 2021.

[14] A. Li, M. Yuan, C. Zheng, and X. Li, "Speech enhancement using progressive learning-based convolutional recurrent neural network," *Appl. Acoust.*, vol. 166, Sep. 2020, Art. no. 107347.

[15] N. Alamdari, A. Azarang, and N. Kehtarnavaz, "Improving deep speech denoising by Noisy2Noisy signal mapping," *Appl. Acoust.*, vol. 172, Jan. 2021, Art. no. 107631.

[16] F. G. Germain, Q. Chen, and V. Koltun, "Speech denoising with deep feature losses," 2018, *arXiv:1806.10522*.

[17] K. Wang, B. He, and W. Zhu, "TSTNN: Two-stage transformer based neural network for speech enhancement in the time domain," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 7098–7102.

[18] Y. Zhang and J. Li, "BirdSoundsDenoising: Deep visual audio denoising for bird sounds," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2023, pp. 2247–2256.

[19] Q. Kong, H. Liu, X. Du, L. Chen, R. Xia, and Y. Wang, "Speech enhancement with weakly labelled data from AudioSet," 2021, *arXiv:2102.09971*.

[20] C. R. Chen, Q. Fan, and R. Panda, "CrossViT: Cross-attention multi-scale vision transformer for image classification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 347–356.

[21] Y. Gong, Y.-A. Chung, and J. Glass, "AST: Audio spectrogram transformer," 2021, *arXiv:2104.01778*.

[22] L. Wang, W. Zheng, X. Ma, and S. Lin, "Denoising speech based on deep learning and wavelet decomposition," *Sci. Program.*, vol. 2021, pp. 1–10, Jul. 2021.

[23] C. Mateo and J. A. Talavera, "Short-time Fourier transform with the window size fixed in the frequency domain," *Digit. Signal Process.*, vol. 77, pp. 13–21, Jun. 2018.

[24] J. Gu, H. Kwon, D. Wang, W. Ye, M. Li, Y. Chen, L. Lai, V. Chandra, and D. Z. Pan, "Multi-scale high-resolution vision transformer for semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 12084–12093.

[25] T.-H. Lin, S.-H. Fang, and Y. Tsao, "Improving biodiversity assessment via unsupervised separation of biological sounds from long-duration recordings," *Sci. Rep.*, vol. 7, no. 1, pp. 1–10, Jul. 2017.

[26] Z. Zha, X. Yuan, B. Wen, J. Zhou, J. Zhang, and C. Zhu, "From rank estimation to rank approximation: Rank residual constraint for image restoration," *IEEE Trans. Image Process.*, vol. 29, pp. 3254–3269, 2020.

[27] Z. Zha, X. Yuan, J. Zhou, C. Zhu, and B. Wen, "Image restoration via simultaneous nonlocal self-similarity priors," *IEEE Trans. Image Process.*, vol. 29, pp. 8561–8576, 2020.

[28] Z. Zha, X. Yuan, B. Wen, J. Zhou, and C. Zhu, "Group sparsity residual constraint with non-local priors for image restoration," *IEEE Trans. Image Process.*, vol. 29, pp. 8960–8975, 2020.

[29] N. S. Haider, R. Periyasamy, D. Joshi, and B. K. Singh, "Savitzky–Golay filter for denoising lung sound," *Brazilian Arch. Biol. Technol.*, vol. 61, pp. 1–10, Nov. 2018.

[30] A. Harma, "Automatic identification of bird species based on sinusoidal modeling of syllables," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, vol. 5, Apr. 2003, p. 545.

[31] B. C. Pijanowski, L. J. Villanueva-Rivera, S. L. Dumyahn, A. Farina, B. L. Krause, B. M. Napoletano, S. H. Gage, and N. Pieretti, "Soundscape ecology: The science of sound in the landscape," *BioScience*, vol. 61, no. 3, pp. 203–216, Mar. 2011.

[32] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 3–19.

[33] H. Hu, Z. Zhang, Z. Xie, and S. Lin, "Local relation networks for image recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3463–3472.

[34] K. Tan and D. Wang, "A convolutional recurrent neural network for real-time speech enhancement," in *Proc. Interspeech*, Sep. 2018, pp. 3229–3233.

[35] Y. Xu, J. Du, L. Dai, and C. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 23, no. 1, pp. 7–19, Jan. 2015.

[36] J. H. L. Hansen, V. Radhakrishnan, and K. H. Arehart, "Speech enhancement based on generalized minimum mean square error estimators and masking properties of the auditory system," *IEEE Trans. Audio, Speech, Language Process.*, vol. 14, no. 6, pp. 2049–2063, Nov. 2006.

[37] S. Sonning, C. Schüldt, H. Erdogan, and S. Wisdom, "Performance study of a convolutional time-domain audio separation network for real-time speech denoising," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 831–835.

[38] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder–decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.

[39] R. Strudel, R. Garcia, I. Laptev, and C. Schmid, "Segmenter: Transformer for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 7242–7252.

[40] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent*. Cham, Switzerland: Springer, 2015, pp. 234–241.
[41] H. Wang, S. Xie, L. Lin, Y. Iwamoto, X. Han, Y. Chen, and R. Tong, "Mixed transformer U-Net for medical image segmentation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2022, pp. 2390–2394.
[42] T. Huang, S. Li, X. Jia, H. Lu, and J. Liu, "Neighbor2Neighbor: Self-supervised denoising from single noisy images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 14776–14785.
[43] A. Sivaraman and M. Kim, "Efficient personalized speech enhancement through self-supervised learning," *IEEE J. Sel. Topics Signal Process.*, vol. 16, no. 6, pp. 1342–1356, Oct. 2022.
[44] W. Yu, J. Zhou, H. Wang, and L. Tao, "SETransformer: Speech enhancement transformer," *Cognit. Comput.*, vol. 14, pp. 1152–1158, May 2022.
[45] T. Fujimura, Y. Koizumi, K. Yatabe, and R. Miyazaki, "Noisy-target training: A training strategy for DNN-based speech enhancement without clean speech," in *Proc. 29th Eur. Signal Process. Conf. (EUSIPCO)*, Aug. 2021, pp. 436–440.
[46] X. Qin, Z. Zhang, C. Huang, M. Dehghan, O. R. Zaiane, and M. Jagersand, "U2-Net: Going deeper with nested U-structure for salient object detection," *Pattern Recognit.*, vol. 106, Oct. 2020, Art. no. 107404.
[47] R. Xu, R. Wu, Y. Ishiwaka, C. Vondrick, and C. Zheng, "Listening to sounds of silence for speech denoising," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 9633–9648.
[48] S. R. Park and J. W. Lee, "A fully convolutional neural network for speech enhancement," in *Proc. Interspeech*, Aug. 2017, pp. 1993–1997.
[49] M. M. Kashyap, A. Tambwekar, K. Manohara, and S. Natarajan, "Speech denoising without clean training data: A Noise2Noise approach," in *Proc. Interspeech*, Aug. 2021, pp. 2716–2720.
[50] E. Moliner and V. Välimäki, "A two-stage U-Net for high-fidelity denoising of historical recordings," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2022, pp. 841–845.
[51] B. Dong, W. Wang, D.-P. Fan, J. Li, H. Fu, and L. Shao, "Polyp-PVT: Polyp segmentation with pyramid vision transformers," 2021, *arXiv:2108.06932*.
[52] W. Wang, E. Xie, X. Li, D. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 548–558.
[53] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9992–10002.
[54] R. Azad, M. Heidari, M. Shariatnia, E. K. Aghdam, S. Karimijafarbigloo, E. Adeli, and D. Merhof, "TransDeepLab: Convolution-free transformer-based DeepLab V3+ for medical image segmentation," in *Proc. Int. Workshop PRedictive Intell. MEdicine*. Cham, Switzerland: Springer, 2022, pp. 91–102.
[55] N. Saleem and M. I. Khattak, "Deep neural networks for speech enhancement in complex-noisy environments," *Int. J. Interact. Multim. Artif. Intell.*, vol. 6, no. 1, pp. 84–90, 2020.
[56] S. R. Park and J. Lee, "A fully convolutional neural network for speech enhancement," 2016, *arXiv:1609.07132*.
[57] N. Priyadarshani, S. Marsland, I. Castro, and A. Punchihewa, "Birdsong denoising using wavelets," *PLoS ONE*, vol. 11, no. 1, Jan. 2016, Art. no. e0146790.
[58] D. Gradolewski and G. Redlarski, "Wavelet-based denoising method for real phonocardiography signal recorded by mobile devices in noisy environment," *Comput. Biol. Med.*, vol. 52, pp. 119–129, Sep. 2014.
[59] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.

JUNHUI LI was born in 1994. He is currently pursuing the master's degree with the Department of Mathematics, School of Science, University of Science and Technology Liaoning (USTL), Anshan, China. His research interests include deep learning, computer vision, and speech enhancement.

PU WANG was born in 2002. She is currently pursuing the bachelor's degree in information and computational science with the School of Science, University of Science and Technology Liaoning (USTL), Anshan, China. Her research interests include deep learning and audio denoising.

YOUSHAN ZHANG (Member, IEEE) received the Ph.D. degree in computer science from Lehigh University. He was a Postdoctoral Researcher with Cornell University. He is currently an Assistant Professor in computer science and artificial intelligence with Yeshiva University. His research interests include artificial intelligence, machine learning, computer vision, transfer learning, manifold learning, and shape analysis.