IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 5 December 2023, accepted 17 December 2023, date of publication 21 December 2023,
date of current version 27 December 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3344813

TOPICAL REVIEW

A Survey of Audio Enhancement Algorithms for Music, Speech, Bioacoustics, Biomedical, Industrial, and Environmental Sounds by Image U-Net

SANIA GUL [1,3] AND MUHAMMAD SALMAN KHAN [2]

[1]Department of Electrical Engineering, University of Engineering and Technology, Peshawar, Peshawar 25000, Pakistan
[2]Department of Electrical Engineering, College of Engineering, Qatar University, Doha, Qatar
[3]Intelligent Information Processing Laboratory, National Center of Artificial Intelligence, University of Engineering and Technology, Peshawar, Peshawar 25000, Pakistan

Corresponding author: Muhammad Salman Khan (salman@qu.edu.qa)

This work was supported by Qatar University under Grant QUST-2-CENG-2023-1640.

ABSTRACT The recent surge in the use of Deep Neural Networks (DNNs) has also made its mark in the field of Audio Enhancement (AE), providing much better quality than the classical methods. Although, there are dedicated audio processing DNNs, yet, many recent models of AE have utilized U-Net: a DNN based on Convolutional Neural Network (CNN), fundamentally developed for image segmentation. It is found that the useful features hidden in the time domain are highlighted when the audio signal is converted to a spectrogram, which can be treated as an image. In this article, we will review the recent work, utilizing U-Nets for different AE applications. Different than other published reviews, this review focuses entirely on AE techniques based on image U-Nets. We will discuss the need for AE, U-Net comparison to other DNNs, the benefits of converting the audio to 2D, input representations that are useful for different AE applications, the architecture of vanilla U-Net and the pre-trained models, variations in vanilla architecture incorporated in different E models, and the state-of-the-art AE algorithms based on U-Net in various applications. Apart from speech and music, this article discusses a wide range of audio signals e.g. environmental, biomedical, bioacoustics, and industrial sounds, not covered collectively in a single article in previously published studies. The article ends with the discussion of colored spectrograms in future AE applications.

INDEX TERMS CNNs, image processing deep neural networks, pre-trained networks, spectrogram, U-Net.

I. INTRODUCTION

Audio Enhancement (AE) is the process of improving the audio quality by removing the noise (produced by the surrounding sources or from the same source in the form of echoes) and filling the gaps due to damage or intrusions [1]. Whether the audio is generated by the objects surrounding us (e.g. people, animals, birds, wind, thunder, traffic, machines, airplanes, musical instruments, etc.) or it is generated artificially by using sophisticated methods (e.g. by [2], or [3] producing audio of high quality and fidelity), before the audio reaches the listener, it is affected by the external factors such as background noise, dereverberation, and competing speakers, badly affecting its quality. This poses several challenges to effective audio communication, especially for machines and hearing-impaired listeners. The AE algorithms ensure to restoration of the quality of audio in the face of such challenges [4]. Thanks to the marvelous capabilities of the human auditory system, people with normal hearing can separate a sound of their interest from multiple simultaneous sounds (cacophony) in a split second. However, people with hearing problems generally face difficulty in doing so.

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

AE is sometimes also required by normal-hearing listeners. To separate the background music from vocals to be reused for karaoke, extract the message from a noisy voice note, recover a classical song from an old damaged gramophone record, separate the contents of the speaker from videos shot at the seashore or busy roads, detect anomaly from the machine sound before the occurrence of fault in the noisy industrial environment, and detecting the endangered species and their number from the sounds recorded in forest are some of the examples, where a normal listener would also require AE. Also, with many new voice-controlled applications emerging, the need to equip machines with robust human-like hearing capability is increasing. Automatic detection of rare events in the dark, automatic fall detection, pathology classification from the auscultation sounds, environmental sounds classification, and source localization for automatic camera maneuver or alarming the rescue workers are a few examples, where machine needs as resilient listening, as humans are rewarded with. However, there is very little probability that the required audio signal is free from the corrupting background noise. Similarly, the efficiency of automatic speech and speaker detection algorithms depends heavily on the quality of input, which is usually deteriorated by the presence of the surrounding noise. So, in all these applications AE is the utmost necessity.

Since its inception in the mid of 1950s, deep learning methods have been responsible for astonishing breakthroughs in every sphere of life including computer vision, speech recognition, natural language processing, bioinformatics, finance and accounting, market predictions, drug design, medical image analysis, climate science, material inspection and gaming [5]. In the past, signal processing and machine learning have been extensively used for AE. However the recent trend of using Deep Neural Networks (DNNs) has generated comparable or even more effective performance than the traditional methods, provided a sufficient amount of training data is available for these networks [6].

The two major types of Deep Neural Networks (DNNs) are i) Recurrent Neural Networks (RNNs) and ii) Feed-Forward Networks (FFNs) (or Multi-Layer Perceptrons (MLPs)), the most popular being the Convolutional Neural Networks (CNNs) [7]. RNNs were developed for processing sequential data, such as text and speech, and have brought significant improvement in speech recognition and natural language processing applications, while, the idea of CNN emerged during Hubel and Wiesel's classic work on the cat's primary visual cortex [8] in 1962. That idea was first put into realization in [9], and later refined in [10]. In both [9] and [10], it was used for image processing, and to date, CNN architecture is enjoying the status of being the most popular network among researchers, for processing the images [11]. Different versions of CNNs were later developed, which can process the 1D time series data (e.g. audio, text, or accelerometer data) directly or on features extracted from sound and the 3D data (e.g. video (sequence of image frames), Magnetic Resonance Imaging (MRI) and Computerized Tomography (CT) scan). The original version of CNN for processing the 2D images (both grayscale and color) is generally called 2D CNN, while the other two versions are called 1D and 3D CNNs respectively. In subsequent discussion, the terms ''image CNN'' and '2D CNN' are used interchangeably highlighting the fact that the initial goal of 2D CNN was image processing.

In recent years, a new version of DNN has evolved, which merges the recurrent layers with the convolutional layers, called Convolutional Recurrent Neural Networks (CRNN) [12]. In CRNN, CNN is used for local feature extraction, while RNN acts as a temporal summarizer, aggregating these features over time to enable the network to take the global structure into account [13]. CRNN, first proposed in [12] for document classification, was later applied to image classification [14] and AE [15].

U-Net is a special CNN-based deep architecture, consisting of many convolutional layers. U-Net was initially proposed for biomedical image segmentation [16] and since then it has become hugely popular within the image/ video processing community. U-Net acquires its name from its architecture, which resembles the shape of an English letter 'U'. Like CNNs, many variants of U-Net also exist, which can even process the audio directly in the time domain (e.g. Wave-U-Net [17], attention Wave-U-Net [18], or Tiny Recurrent U-Net (TRU-Net) used in speech enhancement model of [19]), but we will restrict our discussion to the conventional U-Net, which accept only an image at its input. The difference between an image CNN and an image U-Net is that the image CNN predicts the class of the whole image, while the image U-Net is used for the classification of each pixel of an image. The process is known as ''segmentation''. In this process, the grouping of pixels, belonging to the same class, is done. In image CNNs, the input is an image, while the output is a string of characters (label), defining a unique class for the entire image. In image U-Net, the input is an image, while the output is a matrix of labels defining a unique class for each pixel.

A. RELATED WORK

The most notable reviews of AE using DNNs are [4], [6], [20], and [21] to [28]. While [4], [6], and [20] discuss AE applications restricted to speech and music based on a variety of DNNs, the model in [21] is focused only on Deep Reinforcement Learning (DRL) models covering a wide range of applications including Human-Robot Interaction (HRI), music listening and generation, AE, emotions modeling, spoken dialogue systems and automatic speech recognition. The paper in [22] reviews the DNN-based AE models used exclusively for automatic speech recognition applications. The research [23] also reviews only speech enhancement models using deep diffusion networks and [24] is also dedicated to speech enhancement DNN models. The review [25] again focuses on speech enhancement models using audio-visual deep Kalman filter generative models.

The paper [26] discusses the speech extraction models based on different DNNs using audio, video, spatial, or voice clues of the target. The work in [27] reviews only the source separation application of AE using independent vector analysis. The review in [28] is focused only on machine learning and deep learning models used for AE in hearing aids.

B. OUR CONTRIBUTION

The main contributions of this paper are summarized below.

- In this paper, we give a concise review and insight on AE models, based on image U-Nets. Unlike previously published reviews (e.g. [4], [6], [20], [22], [24], and [26]), which cover all sorts of DNNs, we focus only on image U-Net models used for AE. To the best of our knowledge, this is the first such review of AE, using only 2D U-Nets, is presented.
- Also, as opposed to the earlier reviews, which are mostly restricted to speech and music, our article covers a wide range of acoustic signals, including environmental, biomedical, bioacoustics, and industrial sounds, along with speech and music. Here, we take a comprehensive review of the AE models, based on image U-Nets, for applications including source separation, denoising, dereverberation, and inpainting only.
- In the end, we propose the use of colored spectrograms for AE which although exist for classification AE tasks but a novel idea for the AE tasks requiring image segmentation.

In section II, we compare different audio-processing DNNs with U-Net. In section III we briefly introduce the spectrograms. In section IV, we discuss the need to convert the audio to spectrogram, while in section V the U-Net architecture and the pre-trained audio and image processing models are discussed. Section VI describes the modified architectures and lists their benefits over the vanilla architecture. Section VII describes the input representations commonly used for U-Net-based AE models. In section VIII, a few applications of AE and the State-Of-The-Art (SOTA) models implementing them are described briefly. Section X gives potential directions for future research and the article is concluded in section 10.

II. U-NET COMPARISON TO OTHER DEEP NETWORKS FOR AUDIO ENHANCEMENT

Deep learning has altered dramatically the AE techniques. Given a sufficient amount of training samples, the DNNs have outperformed the traditional signal processing methods, especially under extremely low Signal-to-Noise Ratios (SNRs) and non-stationary noise types [29]. As already pointed out in the discussion above, the DNNs initially developed for directly processing audio signals was Recurrent Neural Network (RNN), but the standard RNN has the problem of taking into account only short-term dependencies due to exploding/ vanishing gradient problems. This problem was resolved by the Long-Short-Term Memory (LSTM) network which uses memory cells to control the flow of information and take care of the long-term dependencies [30]. However, the conventional LSTM is unidirectional and cannot model the future context. To address this issue, bidirectional LSTM networks are introduced which keep in view the future context along with the past context [30]. In 2014 Gated Recurrent Unit (GRU) was introduced for processing sequential data with slight modifications in the LSTM architecture [31]. The computational cost and complexity for all the above-mentioned dedicated end-to-end time domain processing networks dealing with data rates of more than 16 kHz are very large due to the enormous memory requirements to hold the long and short-term dependencies. On the other hand, image processing networks (e.g. Convolutional Neural Networks (CNNs) and image U-Nets) are focused on processing data in a grid-like topology resulting in much reduced computational cost, trainable parameters, and memory requirements than the above-mentioned dedicated audio processing architectures [32]. Since the take-off of the modern deep learning era [33] in 2009 (after Stanford's Fei-Fei Li created ImageNet [34]) many state-of-the-art network architectures and pretrained models have emerged for image processing. Most AE models at that time did not operate directly in the time domain but used the Time-Frequency (TF) domain as inputs and outputs [35]. This 2D audio representation was easily transferrable to image-processing deep learning networks. So, although there exist dedicated audio processing DNNs that can process the signal directly in the time domain, the main motivation for using U-Net (in particular) or other image processing networks (in general) for AE applications was to leverage the extensive research in the fields of images to the field of audio. U-Net; a DNN fundamentally developed for image segmentation has reported high performance when used for AE applications [36]. In the case of natural images, displacement by a single pixel is not perceivable by the human eye. However, in the frequency domain, even a minor shift in the spectrogram has a disastrous effect on listening perception. Similarly, a shift in the time domain is audible as jitter and other artifacts. To preserve the high-level details, the skip connections between the adjacent layers of the same hierarchal level in the encoder and decoder of U-Net play a key role in allowing low-level information to flow directly from the high-resolution input to the high-resolution output [37]. Such connections are not present in auto-encoder and Variational Auto-Encoder (VAE) making U-Net a better choice for AE tasks. In earlier models of AE, only the magnitude spectrogram is enhanced using the noisy phase for reconstruction. The phase is believed not to be corrupted much under high SNR levels but it is highly distorted under low SNR conditions [29]. The introduction of deep complex U-Net has made it possible to incorporate the phase information in the estimated audio, improving the generated signal quality by a large amount over the models using only the magnitude spectrograms [38]. The models having U-Net encoder-decoder architecture and working in

TF complex domain (e.g. [39] and [40]) have been shown to outperform U-Net models working only on TF magnitudes, VAEs, time domain generative models (e.g. SEGAN [41]), time domain TASNet [42] models and its variants (e.g. ConvTasNet [43]), and time domain encoder-decoder models (e.g. Wave-U-Nets [17] and Demucs [44]) at extremely low SNR conditions ranging from -30 to 0dB and non-stationary noise. They offer the best trade between the model size and AE performance when compared to the time domain generative models [29]. Recently diffusion-based generative AE models become popular due to their ability to generalize well with the unseen conditions of noise types, reverberations and SNRs. In pioneer TF-based AE diffusion models U-Net has been an integral part of their structure [23], as will be described in section VIII. All the acoustics applications discussed in this article (separation, inpainting, denoising, and dereverberation) can be merged under the single umbrella of denoising as they all are meant to remove the unwanted sources (whether competitive (in case of source separation), transient (in case of inpainting), diffuse (in case of denoising), or the echoes (in case of dereverberation)) and the AE is meant to restore the audio to as closed to its original generated form as possible.

III. SPECTROGRAMS

Visible sound portrayal is the process of converting the audio signal to an image or portrait. It attempts to perform an analysis similar to that of an ear and presents the result in an orderly manner to the eye [45]. It is believed, that the sound (in time domain) entering the ear is broken down into a collection of local Time-Frequency (TF) regions, before being further processed by the brain [46]). Similarly, in visible sound portrayal, Fourier analysis is performed over small overlapping chunks, and the process is repeated sequentially over a long vector of samples, resulting in a graph, called the Short-Time Fourier Transform (STFT) spectrogram or a standard spectrogram. For audio analysis, the spectrogram is an excellent method of visualizing the signal spectral contents and how they change over time [47]. The final STFT graph has time along the x-axis, frequency along the y-axis, and the brightness or color (along the z-axis) represents the strength of a frequency component at each time frame [48]. In contrast, the standard Fourier transform provides the frequency information averaged over the entire signal interval. The spectrogram was invented in 1940, to help break enemy codes and detect their submarines [45]. Soon, it became a favorite choice for audio signal processing. While there are many visible audio representations, the spectrogram is so common among them, and other representations are visually so much similar to it, that almost every representation of sound in the form of an image, is termed as a 'spectrogram' in literature [48].

Treating the spectrogram as an image, and using the dedicated image DNNs for its enhancement, is an idea first conceived by Humphrey and Bello [49], who classified musical chords in 2012 by an image CNN. Since then many AE techniques, have opted for the use of CNNs (DNNs fundamentally developed for image/ video processing) by converting the audio to a spectrogram and the results are encouraging. The advantage of using the spectrogram-based DNN models for AE includes a lesser number of trainable network parameters and their lesser training cost than the waveform-based models [50]. However, their usage for audio signals requires an additional step of audio signal conversion from 1D to spectrogram, as these networks require input in the form of an image.

IV. NEED OF CONVERSION OF AUDIO FROM 1D TO 2D

In the time domain, once it is mixed, the desired audio signal becomes entirely unidentifiable from the other interfering signals as shown in Figure 1 (top row). So, separating it from noise and reverberations (noise created by the source itself) directly in the time domain is a difficult task. Although there exist such deep learning networks e.g. TASNet [42] and Wave U-Nets [17], which can enhance the signal directly in the time domain, these methods are characterized by their slow convergence, a large number of trainable parameters, and heavy computational load [51].

In the case of speech, surprisingly, the conversion of an audio mixture (consisting of desired signal and noise) from the time domain to the Time-Frequency (TF) domain (i.e. to a spectrogram) makes the target signal discernable, as the simultaneous active speech sources rarely excite the same frequencies at the same time [52]. This rule also applies to the audio mixture of animal sounds [53]. As shown in Figure 1 (top row), once mixed the individual sources S_1 and S_2 are indistinguishable in the time domain. However, in the case of their TF representation, the spectrogram of the audio mixture still has some identifiable portions of S_1 (bounded by red boxes), which can be differentiated from S_2 even by the visual inspection of the mixture's spectrogram. However, this ease of differentiation between sources in the TF domain decreases, if the number of sources or intensity of reverberations increases.

In the case of music, the frequency bands of vocals and different instruments are highly distinctive in the TF domain, as shown by their spectrograms in Figure 2. Percussive instruments, for example, drums have a much flatter spectrum and are well localized in time. In the case of harmonic instruments, e.g. guitar, only a few harmonics are energized at any time, while the piano exhibits both the percussive and the harmonic properties. Vocals, on the other hand, exhibit a higher rate of pitch fluctuation as compared to the instruments [54]. So, based on these distinctive features, vocals or instruments are much more easily extracted from the song (mixture of vocals and instruments) or music (mixture of only instruments) in the TF domain, than in the time domain. Also, the anomalous events, whether in machines, open environments, or biomedical sounds are easily detectable, when converted to the TF domain, as all these applications require a sudden change in acoustic energy

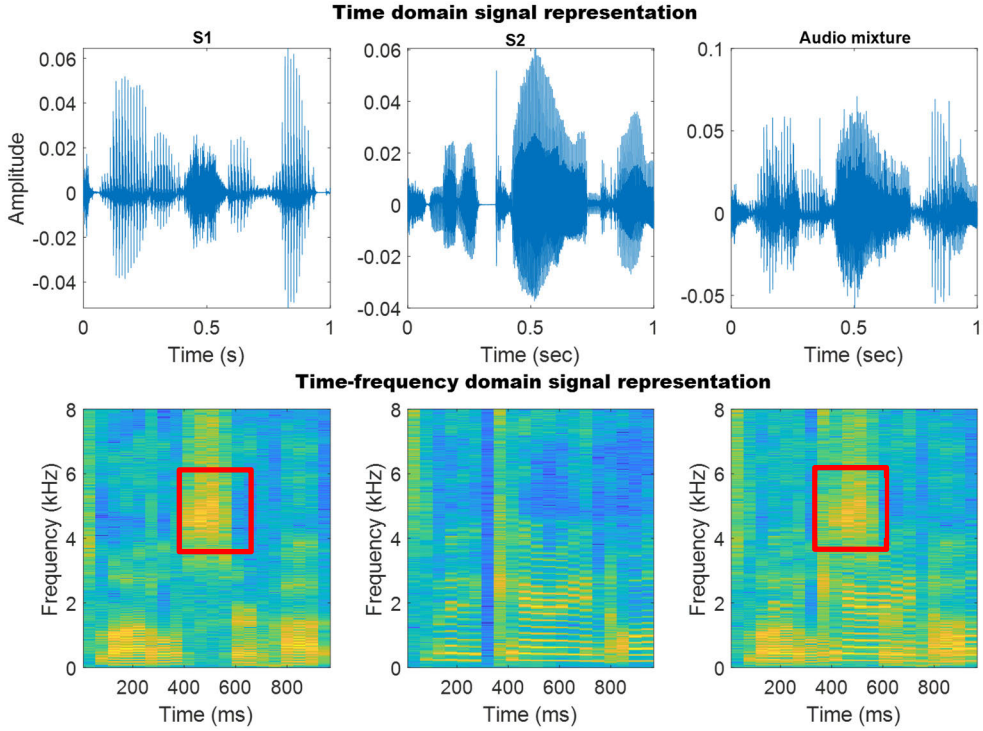

FIGURE 1. Time domain (top row) and time frequency domain (bottom row) representation of speech sources and their resulting audio mixtures. The red boxes show the portions of S1, which are still identifiable in the spectrogram of audio mixture.

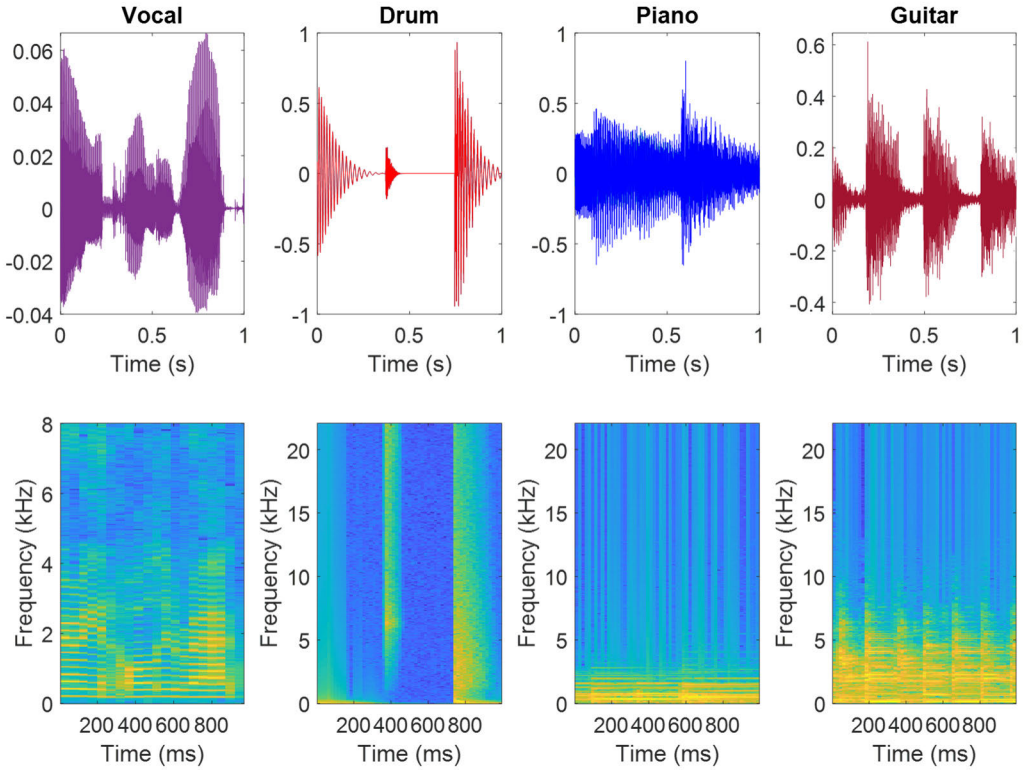

FIGURE 2. Music in time domain (top row) and spectrogram (bottom row) representation.

TABLE 1. Modified architectures of Image U-Net and their advantages.

Completely new architectures	
Deep Complex U-Net	It is a specialized U-Net architecture that uses convolutional layers that can perform convolution operations on complex numbers. The basic reason for their use is to process the phase information along with the magnitude of STFT, as the phase plays a vital role in audio enhancement and recovery [38].
Conditioned U-Net with FiLM (Feature-wIse Linear Modulation) layers	FiLM layers can be inserted at any depth in U-Net and carry out simple feature-wise affine transformations on the neural network's intermediate features based on the input given by the user [70]. The control mechanism is a neural network that generates parameters for the FiLM layers according to the input conditions. Instead of training multiple U-Nets, each for an individual source in the audio mixture, a single network can be used for multiple sources [71]. However, the downside of this network is that it requires multiple passes of a data sample in a single epoch [50].
Multi-channel U-Net	As opposed to the conditioned U-Net the multi-channel U-Net architecture estimates multiple outputs simultaneously from a single network without any conditioning or requiring the training of a single sample multiple times in an epoch [50].
Dense Convolutional Network (DesneNet)	As the convolutional networks become deeper, the information from the input usually washes out due to the vanishing gradient descent problem. To avoid this problem shorter connections are created by connecting the output of each layer to the successive layers creating a dense connectivity pattern called DenseNet [72]. They offer parameter efficiency, easy training, and a regularizing effect reducing the overfitting on smaller datasets.
Attention Gated Control (AGC)-U-Net)	Inspired by the human way of handling massive information collected by the eyes and the ears, focusing on important features and discarding the irrelevant ones [73], the attention mechanism added on the skip connections in AGC-U-Net architecture discards the noise and reduces the semantic gap between the low-level features on the encoding side and the high-level features on the decoding side [74].
U-Net++ [75]	To fill the semantic gap between the encoder and the decoder of vanilla architecture, skip connections are redesigned through a series of nested dense convolutional blocks. Making the feature maps similar at the encoder and decoder side results in making the learning task easier for the optimizer.
Limited Upscale U-Net (LUU-Net)	In vanilla U-Net, both time and frequency axes are restored to their original dimensions at the output of the decoder side. However, the event detection tasks do not require the frequency axis restoration as the required information lies on the time axis. So in LUU-Net only time scale is restored by limited upscaling of the frequency axis by using asymmetric stride for decoder convolutional layers [68]. This results in an immense reduction in learnable parameters.
U^2-Net	It is a two-level nested U-Net structure [76], which provides high resolution without increasing the computational and memory costs as compared to the vanilla architecture.
Hyper-parameter modifications in the vanilla architecture	
Dilated convolution layers	The dilated convolution layers use dilated filters. This filter expands the input by setting holes between its consecutive elements [40]. This results in enlarging the receptive field and thus enables it to find long-term dependencies in the input without increasing the number of parameters [40].
Leaky Integrate and Fire (LIF) activation function	Instead of the continuous activations as present in the conventional DNNs, the LIF activation is discrete which offers simplicity and computational efficiency [77].
Asymmetric filters	As the pitch usually occupies several frequency bands, the filter must be longer on the frequency dimension than in the time dimension to better capture the spectral patterns of speech [66]. Using U-Net for ordinary images, the convolution filters are usually symmetric but for speech asymmetric filters perform better [66].
Transformations in the bottleneck	
Variational Auto-Encoder (VAE)	In contrast to the deterministic characteristics of vanilla U-Net, the VAE in the bottleneck offers increased robustness towards out-of-distribution effects, such as reverberation and unknown noise types [65].
Bidirectional Long Short-Term Memory (BLSTM)	The BLSTM in the bottleneck ensures the extraction of long-term temporal information present in audio.
Cross-modal early fusion	This layer in the bottleneck concatenates the audio and video (weighted by attention matrix) features [78].
Intermediate layers on the encoding and decoding sides	
Time-distributed and time-frequency-distributed blocks	The time-distributed blocks are used to extract the long-range correlations that exist along the frequency axis and the time-frequency-distributed blocks are used to extract them along both the time and frequency axis of the spectrogram [39].
Convolution Attention (CA) blocks	In CA blocks time-attention mechanisms are combined with sequential convolutions to learn both global and local dependencies [79].
Recurrent-Neural-Network (RNN) Attention (RA) blocks and the Res Paths (RPs)	The RA blocks in the skip connections increase the effective receptive field and explore the most efficient representations with frequency-specific characteristics, while the presence of RP blocks avoids immediate integration of low-level features on the encoder side with the high-level features on the decoder side by first reducing the semantic gap between them [80].

TABLE 1. *(Continued.)* **Modified architectures of Image U-Net and their advantages.**

Residual Blocks (RB)	When shift-based operations are performed on complex numbers in convolutional layers, imbalances are introduced between real and imaginary components, leading to perceptual artifacts in the generated output. RB solves this issue [81].
Multi-Lane Dimensionality Reduction (MLDR) module	MLDR module performs dimensionality reduction between 2D convolutional processes. It reduces the number of trainable parameters through factorization of the multi-dimensional filter operation [82].

which is performed better in the TF domain than in the time domain [55]. It is also found to be beneficial for the recovery of missing audio, in case of loss by sudden intrusions.

V. U-NET ARCHITECTURE AND THE PRE-TRAINED NETWORKS

U-Net architecture contains two paths, an encoder and a symmetric decoder. In its basic architecture, the encoder side is composed of multiple sets of three types of layers i.e. i) convolutional, ii) nonlinear, and iii) the pooling layer [56]. The encoder path is the contracting path; where the convolutional layers extract the features starting from the very basic level e.g. edges and corners and continuing with the more abstract ones, as the image moves down the path [57]. The nonlinear layer (activation layer) is responsible for saturating or limiting the generated output of the convolutional layer. Although many types of activation layers exist, the most common ones are sigmoid, Rectified Linear unit (ReLu), leaky ReLu, softmax, Scaled exponential Linear unit (SeLu), and tanh. As the image goes down the encoder path, it is down-sampled by the pooling layers, which in turn makes the computation faster by retaining the important features and dropping the redundant (or nonuseful) ones [56]. The opposite occurs in the decoding path, where the output of the encoding path is upscaled gradually by using the transposed convolution and upsampling layers instead of the pooling layers, till it reaches the size of the input image at the end of the decoding path. Drop-out and batch normalization layers can be added in both paths to avoid the overfitting problem and to attain training stability respectively [55]. However, as the network depth increases, more information gets lost, dropping to its minimum at the bottleneck of U-Net, making it nearly impossible for the decoder to reproduce the image with fine-grained details. So, to overcome this problem, skip connections are provided between the peer layers of the encoder and decoder sides. They alleviate the problem of information loss by bypassing the bottleneck and providing the decoder with the encoder's side high-resolution, fine-grained details. Removing the skip connections would result in the creation of an auto-encoder; a DNN in itself. Although the conventional U-Net has the architecture described above, now more types of DNN layers are being added to make U-Net more adaptable to audio needs.

Like CNN, U-Net is also a supervised neural network, requiring Ground Truth (GT) (annotation / label) for each pixel of the training image. GT is an 'ideal' result, we would expect from our model to predict for a given pixel. During training, the neural network matches its output with the GT to adjust its parameters to achieve prediction accuracy. Like all DNNs, the more data the U-Net is trained over, the more it would generalize and maintain good performance under unseen conditions. In the case of speech, large datasets are present in English, while Western music also enjoys large reservoirs. For other languages and Eastern music, there is a limited reservoir of stored examples. Similarly, in the case of environmental sounds, the size of the available dataset is very small compared to the diversity of these sounds. Artificially generating data and data augmentation: i.e. slightly modifying the existing data e.g. by changing its pitch, stretching in time, or spectral filtering reinforces the otherwise smaller datasets [6].

Apart from the requirement of large datasets, training any DNN is a computationally expensive and lengthy process. In such cases, pre-trained networks already trained over large datasets, are very useful, as they require far less data, time, and computational resources than needed if the system is trained from scratch. Using a pre-trained model on a new problem is called transfer learning. Notable examples of pre-trained image CNNs are AlexNet, DenseNet, GoogLeNet, VGG-16, ResNet, and ZFNet. These models can be used either for image classification in their default architecture or for image segmentation, by replacing their output layers with the decoder of U-Net.

The use of pre-trained networks for audio applications started back in 2014 when Gwardys and Grzywczak [58] used a pre-trained image DNN (an image CNN, which was trained on a dataset with more than one million images; the winner of the Large Scale Visual Recognition Challenge (ILSVRC) 2012) for music genre classification. Although pre-trained networks are available for transfer learning in computer vision problems, there are only a few such networks available in the audio domain. One such network is speechVGG [19], which adopted its architecture from image VGG and is trained on large datasets of spectrograms of the most frequently used words taken from the LibriSpeech dataset [59]. It can be used for transfer learning in applications such as speech inpainting, language identification, speech, noise and music classification, and speaker identification or for estimating the training loss for other DNNs. Other examples of audio pre-trained models are YAMNet (Yet Another Mobile Network; an audio classification network by Google) [47], TRILL

(TRIpLet Loss network) [60], trained on Audioset [61] and BOYLE-A (Bootstrap Your Own Latent for Audio) [62], trained on Audioset [61] and FSD50K [63]. Training on these large datasets enables these networks to learn the distinguishing features of a wide range of human and animal sounds, musical instruments and genres, and common everyday environmental sounds [61]. Among all these pre-trained networks, only speechVGG is trained and accepts the input in the form of a spectrogram, while YAMNet, TRILL, and BOYLE-A are trained on features extracted from the audio.

VI. MODIFIED ARCHITECTURES

As found in many audio enhancement applications the vanilla architecture usually does not generate an impressive audio quality (although comparable or slightly better than the machine learning algorithms e.g. [64], [65], [66], [67], [68], and [69]) in the output. Therefore either additional layers are added, skip connections are modified, hyper parameters are adjusted or completely new architectures are proposed in many recent AE models to improve the performance. Here these modifications and their benefits are listed in Table 1, while their complete details can be obtained from their respective papers.

VII. INPUT REPRESENTATIONS FOR U-NET

For U-Net, the input must be in the form of an image (grayscale or color) [83]. As the image itself is a tensor with a 2D shape and a varying number of channels (1 for grayscale, 3 for colored), any tensor of numbers with the dimensions of an image can act as an input for U-Net, no matter it is visible to the human eye as an ordinary image or not. Our discussion in the next section will be restricted to those SOTA models that can accept data only in the form of images. U-Net-based AE models, which accept the audio itself or features extracted from it in the form of a 1D signal are out of the scope of this paper.

As already described, the most widely used image representation for AE is the standard STFT spectrogram. Although it is the simplest TF representation, it has a few shortcomings including its failure to provide effective resolution for the wide band signals e.g. speech and music [59], insufficient frequency resolution (which prevents algorithms from separating closely spaced tones required for the effective noise removal, the possibility of creating unwanted modulations in channel extraction or time-stretching applications), the introduction of artifacts resulting in pre-echoes (audible as the swishy, non-focused sounding of transients of percussive instruments e.g. drum) and its purely linear frequency resolution, which is not compatible with the human perception [84] as at higher frequencies, human beings perceive sounds logarithmically, rather than linearly. A melodic (Mel) spectrogram plots the frequency on a logarithmic scale, and so conforms well to human perception. The term 'log-scaled' spectrogram, commonly used in literature is not the same as the Mel spectrogram. In the Mel spectrogram, log scaling of the frequency (y) axis is done, while in a log-scaled spectrogram, the brightness (z) axis is log-scaled. Both STFT and Mel spectrograms can be log-scaled and are in common use, apart from their non-log-scaled counterparts [85].

In spectrograms, there is a tradeoff involved in time and frequency resolution, defined by the analysis window size, which is defined as the smallest segment (chunk) of an audio signal over which the Fourier transformation is applied. This window keeps sliding over the upcoming samples and provides time-localized frequency information. An increase in the window size would increase the frequency resolution and reduce the time resolution and vice versa [86]. The window size is chosen by convention for most applications and once chosen, is not considered further. A one-size-fits-all approach does not make sense and sounds from different sources or for different applications usually require tailored window sizes [87]. The appropriate window size can either be selected manually (e.g. based on some prior information about the signal energy profile [87] or the distance between the two closest sinusoids [59]), or by an adaptive mean (e.g. based on some local characteristics of the signal [88]), which requires no prior information about the input signal. However, the adaptive selection method is computationally expensive, as compared to the fixed-size window method.

Apart from STFT and Mel spectrograms, various other 2D representations, e.g. cochleagram, Constant Q Transform (CQT), chromagram, tempogram, Auditory Image Map (AIM), Stabilized Auditory Image (SAI), etc. are in use [48] and many others from the TF gallery [89] e.g. Wigner-Ville distribution, Empirical Mode Decomposition (EMD), Hilbert-Huang Transform (HHT), Fourier Synchro-Squeezed Transform (FSST) can be used in future AE applications to decompose the audio into TF domain to highlight the components of interest in the audio signal. Every representation has its own method of calculation and frequency resolution. As compared to the standard STFT, all other TF representations are computationally expensive, but the shortcomings of STFT (already mentioned above), and the capability of highlighting the desired audio features by others, favor their use [59]. The preference of any particular representation for an application depends on its ability to highlight the discriminatory features in the signal of interest. Some features are readily visible in an ordinary spectrogram (fixed-resolution/ standard STFT). For example, in the case of overlapping events, the properties of each event are more easily identifiable in multi-resolution STFT, while they are diluted in the case of Mel-spectrograms, while the CQT images are useful for music analysis ([48] and [49]). Similarly, few sounds like audio anomalies in machines or open environments are better captured with Mel-spectrograms. This is because, at low SNRs, they successfully highlight the transition of sound generated by the occurrence of an anomalous event [90], while the medical anomalies are usually more easily identified by scalograms [91], which depict better, the slowly varying signals, punctured

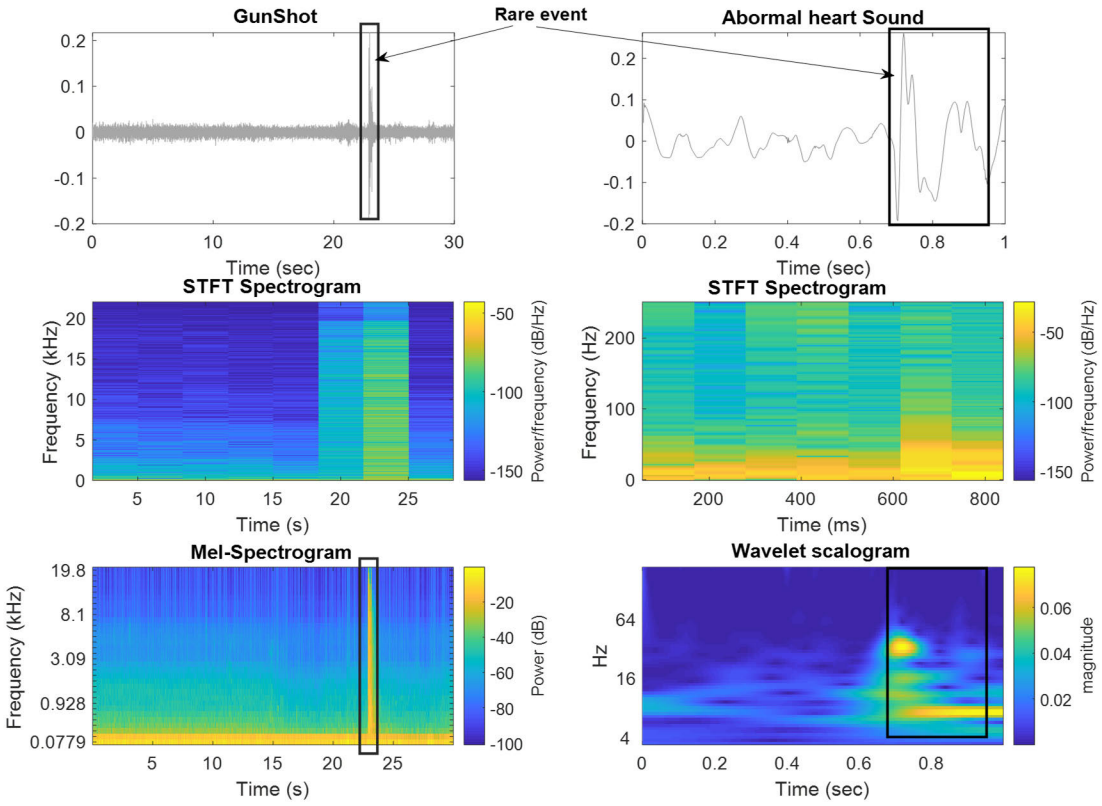

FIGURE 3. Rare sound event detection. STFT spectrogram fails to capture the rare event precisely, while the Mel spectrogram (left) and scalogram (right) have done so for the open environment and health anomaly sounds respectively, as shown by the black boxes bounding the rare events. The STFT and Mel-Spectrograms are plotted by the MATLAB instructions spectrogram(audio,[],[],[],Fs,'yaxis') [253], and melSpectrogram(audio,Fs) [254] respectively, where Fs is the sampling frequency. The helperPlotScalogram function [255] is used for plotting the wavelet scalogram.

by the abrupt transients. The effectiveness of using these representations for detecting anomalies (open environment and medical pathological sounds) as compared to the ordinary STFT is depicted in Figure 3.

In treating the spectrogram as an image, the question arises, whether there exists any similarity between the two. The answer is both yes and no. Like natural images, the spectrograms of natural sounds have a built-in correlation between the neighboring bins [6], but there exist additional correlations at harmonics, which are not found in the case of ordinary images. Furthermore, in contrast to images, the energy distribution differs significantly between frequency bands. This effect is countered by standardizing the spectrogram separately for each band [6].

VIII. AUDIO ENHANCEMENT BY SEGMENTATION

2017 marks the beginning of using U-Net (a specialized DNN for image segmentation) for AE when Jansson and Humphrey [37] used it for separating the vocals from instruments in a song. We believe that we were the first to use U-Net for speech separation [64]. The authors of [92] were the pioneers of using pre-trained model for speech inpainting. Although the model of [93] has used U-Net as the generator of Generative Adversarial Network (GAN) for the first time in speech denoising application, it was in the model [66], that the U-Net was tested for the first time as a solo network for speech denoising and dereverberation.

Common AE applications using image U-Net include audio/ music generation, text-to-audio generation, bandwidth extension, source localization, vocoders, pitch marking, source separation, source inpainting, and source denoising/ dereverberation. However, we will restrict our discussion to only the last three AE applications as depicted in Figure 4. In this paper we will review SOTA models employing image U-Nets for different AE tasks from 2017 (the beginning of U-Net for AE applications) to 2023. This paper is not intended to compare the performance of different models, as they are rarely trained and tested under similar acoustic conditions (reverberation, SNRs, and noise types) and over similar datasets. Neither the evaluation metrics are common among them, except for the Music Source Separation (MSS) task where most of the models use the metrics of [94] for performance evaluation. We will only compare these models in Table 2. For the rest of the tasks discussed in this article, only the salient features including the architecture, the input representation, and the dataset used for training and testing are mentioned in Tables 3, 4, and 5, as reported in their corresponding papers. The

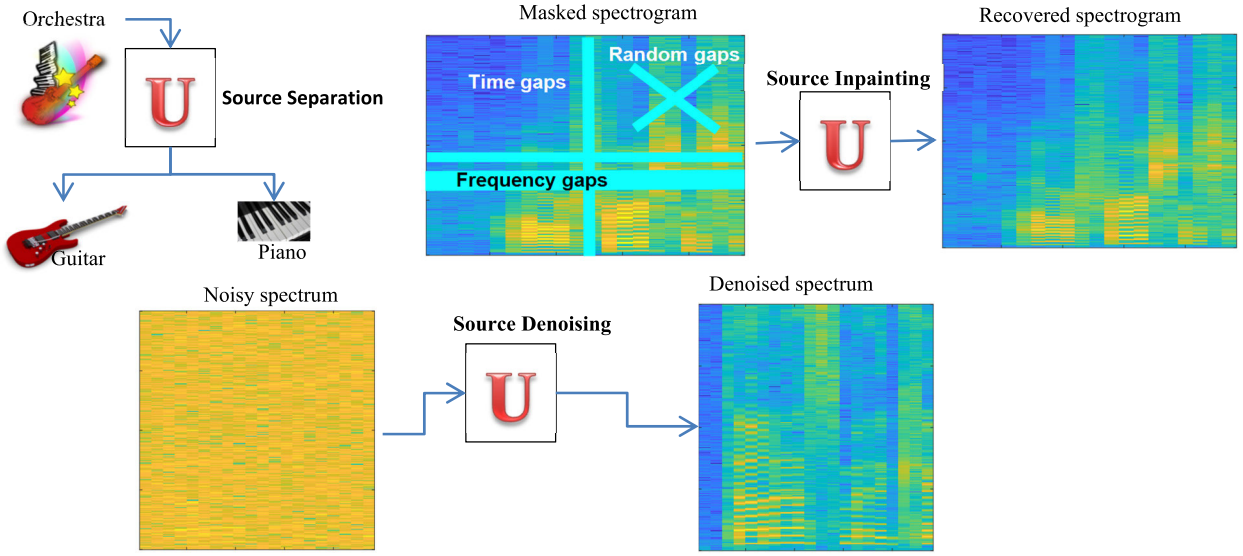

FIGURE 4. Audio enhancement by U-Net.

TABLE 2. Comparison of different MSS models on common datasets.

Model	Musdb18 [103]			Model	MUSIC [104]			Model	URMP [105]		
	SDR↑	SIR↑	SAR↑		SDR↑	SIR↑	SAR↑		SDR↑	SIR↑	SAR↑
[71]	2.4	**7.52**	5.69	[98]	NA	15.02	12.28	[78]	NA	5.41	**11.47**
[50]	3.66	NA	NA	[78]	7.26	14	NA	[100]	**3.05**	**8.5**	7.54
[39]	**7.12**	NA	NA	[100]	5.05	10.93	9.93				
[96]	5.2	6.4	**11.98**	[101]	4.26	7.07	13				
[98]	6.45			[74]	**10.96**	**17.91**	12.77				
				[79]	5.5	7.5	**22.76**				
				[102]	5	NA	5.1				

*NA= Not Available

preference is given to the pioneering and the most recent research.

A. SOURCE SEPARATION (SS)

We divide the Source Separation (SS) problem into three main streams; a) Music Source Separation (MSS), b) Zoological Source Separation (ZSS), and c) Speech Source Separation (SSS). In most of the spectrogram-based methods, the networks are trained either to estimate mappings to a clean target spectrogram or to estimate masks (binary or ratio (real or complex)), that classify every pixel of the spectrogram [95]. These masks are then multiplied with the spectrogram of the noisy mixture, to obtain the estimates of the corresponding sources [50].

1) MUSIC SOURCE SEPARATION (MSS)

The goal of MSS is to design algorithms that can separate vocals from instruments (called Singing Voice Separation (SVS)) or the instruments themselves. U-Net based algorithms provide promising performance for both SVS and MSS in general [39]. These algorithms have been successfully used for music editing/remixing, music retrieval for karaoke, singer identification and transcription of musical records. In case of music instrument separation, CQT has shown superior performance, as compared to other types of representations [6], as shown on the left side of Figure 5, where the two instruments (drum and piano) are visibly more separable in CQT representation than in any other. However, for SVS, the use of standard STFT spectrogram is found to be beneficial ([85] and [96]).

a: AUDIO-BASED MSS

In [37], two U-Nets are used for SVS, which are trained on the STFT spectrograms of the full song, along with the separate STFT spectrograms of vocals and instruments, which act as GT for each network. As already pointed out in the discussion above it was the first DNN model, which has used U-Net for AE, and provided a better quality output signal, with fewer distortions and artifacts, when compared to other deep learning models of MSS. The model of [71], in contrast to [37], consists of a single U-Net trained on the STFT spectrograms, multiple times in every training epoch; each time with a source-specific conditioning. On the other hand, the model of MSS [50] is a multi-task model, using the STFT representation. However, unlike [37], only a single network is trained for all sources, so its training parameters are equal

TABLE 3. Source separating models.

Model	Application	Pre-trained model	Network architecture	Input representation	Dataset used
Music source separation (MSS)					
A. Jansson et al. [37]	Vocal/instrument separation	×	Vanilla U-Net	STFT magnitude spectrogram	1). MedleyDB [117] 2). iKala [118]
G. M. Brocal et al. [71]	Vocals/Drum/bass/Rest separation	×	Conditioned U-Net and Feature-wise Linear Modulation (FiLM) layers inserted the encoder side	STFT magnitude spectrogram	Musdb18 [103]
V. S. Kadandale et al. [50]	1). Vocal/instrument separation 2). Vocals/Drum/bass/Rest separation	×	Multi-channel U-Net	STFT magnitude spectrogram	Musdb18 [103]
W. Choi et al. [39]	Singing voice separation	×	Vanilla U-Net with intermediate layers at both sides consisting of either 1) Time distributed blocks of PHASEN [146] or Time-frequency distributed convolutions.	Complex STFT spectrogram	Musdb18 [103]
A. C. Hadria et al. [96]	Singing voice separation	×	Vanilla U-Net	STFT magnitude spectrogram	Musdb18 [103]
V. George et al. [98]	Drum separation	×	Vanilla U-Net with dense block in the bottleneck	STFT magnitude spectrogram	1). Demixing Secret Data (DSD) [119] 2). Musdb18 [103]
H. Zhao et al. [99]	Instrument separation	ResNet for video	Vanilla U-Net	STFT magnitude spectrogram	MUSIC [104]
C. Gan et al. [78]	Instrument separation	×	Vanilla U-Net with cross-modal early fusion module in the bottleneck	STFT magnitude spectrogram	1). MUSIC [104] 2). URMP [105] 3). AtinPiano [120]
H. Zhao et al. [100]	Instrument separation	×	Vanilla U-Net with Feature-wise Linear Modulation (FiLM) layers inserted in the bottleneck	STFT magnitude spectrogram	1). MUSIC [104] 2). URMP [105]
R. Gao et al. [101]	Instrument separation	ResNet for video	Vanilla U-Net	STFT magnitude spectrogram	1).MUSIC [104] 2). AudioSet [61] 3). AV-Bench [121]
Y. Zhang et al. [74]	Instrument separation	ResNet for video	Attention Gate Control (AGC-U-Net) at the skip connections	Mel-spectrogram	MUSIC [104]
C. Huang et al. [79]	Instrument separation	ResNet for video	Vanilla U-Net with Convolution-Attention (CA) blocks in bottleneck	STFT magnitude spectrogram	1). MUSIC [104] 2). Audio-Visual Event (AVE) dataset [122]
S. Mo et al. [102]	Instrument separation	ResNet for video	Vanilla U-Net	Log Mel spectrogram	1). MUSIC [104] 2). VGG-Instruments [123] 3). VGGMusic [124] 4). VGGSound [124] 5). Kinetics-400 [125]

TABLE 3. *(Continued.)* **Source separating models.**

Zoological sounds separation (ZSS)					
P. C. Bermant et al. [53]	Rhesus macaques, bottlenose dolphins, and Egyptian fruit bats separation and classification	×	Vanilla U-Net	1). Log-scaled STFT spectrogram 2). STFT spectrogram	1). Macaque coo call dataset [126]. 2). Bottlenose dolphin signature whistle dataset [127]. 3). Egyptian fruit bat vocalization dataset [128].
C. Bergler et al. [107]	Killer whale sound separation	×	Vanilla U-Net	Log-scaled STFT spectrogram	Orchive [129]
T. Colligan et al. [108]	Beetle courtships and whale songs classification	×	Vanilla U-Net with 2D convolutional layers replaced by 1D	Mel spectrogram	1). Real recordings of beetle songs 2) Whale detection challenge 2013 [130]
Speech source separation (SSS)					
S. Gul et al. [64]	Speech separation	×	Vanilla U-Net	Spatial spectrograms	1). TIMIT [131] 2). McGill [132]
S. Gul et al. [108]	Speech separation	SONET [64]	Vanilla U-Net	Spatial spectrograms	TIMIT [131]
S. Basir et al. [110]	Speech separation	×	Vanilla U-Net	STFT complex spectrograms	TIMIT [131]
C. Pang et al. [111]	Multi-channel speech separation	×	Vanilla U-Net with dilated convolutions	Complex STFT spectrograms	1). VOICES [133] 2). CHiME-3 [134] 3). WMIR [135]
R. Gao et al. [112]	Audio-visual speech separation	ResNet for video ShuffleNet for audio	Vanilla U-Net	Complex STFT spectrograms	1). VoxCeleb2 [136] 2). Mandarin [137] 3). TCD-TIMIT [138] 4). CUAVE [139] 5). LRS23 [140] 6). VoxCeleb1 [141]
Y. Wu et al. [113]	Audio-visual speech separation	ResNet for video ShuffleNet for audio	Vanilla U-Net	Complex STFT spectrograms	1). VoxCeleb2 [136] 2). Mandarin [137] 3). TCD-TIMIT [138] 4). CUAVE [139] 5). LRS23 [140] 6). VoxCeleb1 [141]
M. Yoshida et al. [114]	Audio-visual speech separation	ResNet for video	Complex U-Net	Complex STFT spectrograms	Fair-Play dataset [142]
G. Dahy et al. [40]	Audio-visual speech separation	×	Vanilla U-Net	Complex STFT spectrograms	BBC (LRS2) Dataset [143]
J.W. Hwang et al. [80]	Audio-visual speech separation & denoising	×	Vanilla U-Net with skip connections having RA and RP blocks	Complex STFT spectrogram	1). LRS2-BBC dataset [143] 2). Voice Bank corpus [144] 3). DEMAND [145]

to a model trained for a single source. Also, unlike [71], training the model multiple times in every epoch is not required, for multiple sources. This model produces output quality comparable to the dedicated and source-conditioned models, with much less computational resources. Although STFT itself is a complex matrix, unfortunately in most SS models, the complex-valued phase is often neglected, due to difficulty in its estimation and the SS models usually estimate magnitude masks, while reusing the noisy phase information of audio mixture for the source retrieval on the assumption that the phase is not highly affected by noise [95]. This has clear limitations, especially under low SNR conditions, where the difference between the clean and the estimated target signals gets larger with decreasing SNR values [97]. The SVS model proposed in [39] uses deep complex U-Net and complex-valued STFT spectrograms, to estimate the

complex-valued Ratio Mask (cRM). As the DNN-based models other than the pre-trained models require a lot of training data, in [96] different augmentation techniques are tested for the U-Net model designed for SVS. It is found that the most effective augmentation technique for U-Net is pitch shifting as compared to time stretching and formant shifting. Few models also incorporate additional layers in the conventional structure of U-Net to achieve better separation e.g. the model in [98], dedicated to separating only the drum sound from the polyphonic music mixture uses a dense block at the bottleneck of U-Net. The dense block has cascaded convolutional layers with each layer connected to all the layers in front of it.

b: AUDIOVISUAL-BASED MSS

The natural synchronization that exists between vision and sound provides a rich supervisory signal to localize and separate the sounds in videos [99]. There are various MSS models (e.g. [78], [99], [100], and [101]) using video in addition to audio. These models use different networks for processing the video while they all use U-Net for processing the audio signal and accept the STFT spectrogram of the audio mixture at their input. The model proposed in [99] is useful for separating a mixture of two instruments. It uses U-Net for audio separation and ResNet for video processing. The model proposed in [78] uses the features extracted from the body movements and processes them by Graph Convolutional Network (GCN). The output of GCN is fused to the middle part of U-Net for guiding the MSS. This method proves very useful, especially for solving the harder problem of homo-musical separation, where two or more people are playing the same instrument in a single frame. The model proposed in [100] incorporates motion cues extracted from a deep dense trajectory network and injects them in the middle of U-Net. This model is based on the observation that if two people are playing the same instrument simultaneously, the humans can separate their beats by observing the movement of each person for a while. The MSS model proposed in [101] is similar in structure to the model of [99] except that it uses ResNet for processing video and injects the features into the middle of U-Net. This model proves beneficial for separating the sounds generated by similar-looking instruments in an unlabeled video. The model in [74] also uses ResNet for video feature extraction, but it merges them with the output of U-Net instead of injecting them in the bottleneck of U-Net. While the above-stated audio-visual MSS models are masking-based, the model proposed in [79] is a generative model that produces higher-quality source separation as compared to the masking-based models. The local pattern from the visual cues is extracted by ResNet and the long-range time dependencies in audio are extracted by the time-attention module and injected into U-Net which primarily does the audio separation task. The unified audiovisual MSS model for localization, separation, and recognition [102] again uses conventional U-Net. The video and location features of different sources collected by DNNs are injected in the middle part of U-Net to guide the audio separation process.

The performance comparison of different MSS models using the evaluation metric of [94] (Signal-to-Distortion Ratio (SDR), Signal-to-Interference Ratio (SIR), and Signal-to-Artifact Ratio (SAR) is given in Table 2. For all metrics, higher is better. Only the models trained and tested on a common dataset are compared with each other. The results are reported directly from their papers. In any paper if separate results have been given for all instruments and vocals, the average of them is given in Table 2. In the source separation task, among all the three metrics of [94], SDR is the most important parameter showing the overall degree of separation. It is the ratio of the signal energy with the sum of interference, artifacts, and noise energies. As clear from Table 2, for the Musdb18 [103] dataset, the model of [39] using the phase information in the Time-Frequency distributions results in the output quality enhancement, followed by the model [98] having a dense block in its bottleneck. For the MUSIC [104] dataset, the model of [74] reducing the semantic gap and introducing the attention mechanism by AGC-U-Net has resulted in the overall best performance than the competitive networks. For the URMP [105] dataset, SDR is not reported for [78]. The model of [78] (audio-visual model) offers better SAR and the model of [100] has better SIR.

In short, the models using phase information, modified architectures with the reduced semantic gap between encoder and decoder, and complemented by visual modality. work better than the audio-only models using the noisy phase and vanilla architecture (e.g. [96], [99], and [101]).

2) ZOOLOGICAL SOUNDS SEPARATION(ZSS)

Due to lesser research in the bio-acoustic field, it remains unclear which input representation is the most suitable one for the machine learning bioacoustics applications [53]. The choice of useful representation for bioacoustics signals may vary according to the vocal properties of the particular species, yet HHT and CWT are optimal representations for detecting whale screams and bird songs respectively [89].

BioCPPNet [53] is a U-Net-based model for ZSS; across diverse biological taxa. The separation results show that the ordinary STFT is more useful than its log-scaled counterpart. This model trains multiple U-Nets, each on an individual species, and after training uses them for ZSS. ORCA-Party Problem (OPP) (a term coined for the largest member of the dolphin family – the killer whale (Orcinus Orca); akin to the term Cocktail Party Problem (CPP) [106], used generally for all types of SS) is solved in [97] by using the conventional U-Net architecture of [16]. It takes the log-scaled STFT spectrograms as input and generates eight different masks for eight different kinds of ORCA sound activities. The model of [107] is a generalized model for the classification of different animal sounds designed originally for beetle courtship sounds classification and later tested for whale song

classification. The model uses an ensemble of U-Nets for noise removal and classification of sounds is later done by computing the median softmax value over all the members in the ensemble.

3) SPEECH SOURCE SEPARATION (SSS)

a: AUDIO-BASED SSS

U-Net is used for binaural SSS in [64] by clustering the TF units of the spatial spectrogram of the audio mixture. The spatial spectrogram is a special spectrogram that contains information about the spatial location of sources, contributing to the audio mixture. This spectrogram is obtained by the ratio of STFT of audio mixtures collected at the two ears. Although this SSS model called 'SONET' was the first of its kind, incorporating U-Net for audio SSS and spatial cues, it fails to generate any noticeable improvement in speech quality over the SOTA spatial cue-based machine learning models. Also, this model is restricted for anechoic conditions. This problem was resolved in [108] by using SONET with Expectation Maximization (EM) (a machine learning algorithm), which outperforms its constituent systems, both under anechoic and reverberant conditions, as indicated by the results of subjective listening tests in [109]. The most interesting fact about the SSS model in [108] is that it uses the anechoic pre-trained model 'SONET', without any need for retraining, to tackle the echoes. SONET is also used for speech dereverberation, as will be discussed in the 'source denoising and deteverberation' subsection. The SSS model proposed in [110] concatenates the real and imaginary components of the STFT matrix which is then processed by ordinary U-Net instead of deep complex U-Net. This model is designed to separate the mixture of male and female speech only. The SSS model [111] is a multichannel speech enhancement model that utilizes beamforming at the frontend to discard the distractors and U-Net to produce separate amplitude and phase spectrograms for each channel. The output of U-Net is given to a post-filtering network which captures contextual and spatial correlation information and generates an estimated spectrogram.

b: AUDIO-VISUAL-BASED SSS

The audio-visual models utilizing U-Net also exist in the case of SSS as there exists a strong link between how a person's face looks and how his voice sounds [112]. The most well-known contribution in this direction using U-Net for SSS is VisualVoice [112], where ResNet extracts the facial attributes from the image and ShuffleNet extracts the features from the lip movement in a video and injects them in the bottleneck of the U-Net performing the SSS task. The model shows superior performance over audio-only speech separation and audio-visual source separation models using DNNs other than U-Net but its major drawback is its high computational cost which is reduced by [113] by incorporating various quantization techniques on the network parameters. The audio-visual SSS model [114] is an interesting model that separates the off-screen sounds (whose sources are outside the video) from the on-screen sources using the interaural cues. The audio collected from the binaural setup is decomposed into left and right spectrograms, and each spectrum is again split into real and imaginary components resulting in four matrices which are stacked together and input into the U-Net. The visual cues are collected from the ResNet18 pre-trained model and concatenated with the spatial cues of audio collected from the S&E model [115] and injected in the bottleneck of U-Net to guide the separation process. The SSS model of [40] is also an audio-visual model, using deep complex U-Net for complex STFT spectrogram processing. However, it requires the training of three different DNNs to separate the two speakers. The first is a pre-trained DNN called FACENET [116] for learning the facial features of the target speaker obtained from the video stream, the second is an RNN for learning the features of the target and masker for the audio stream and the third is U-Net for learning the spectral features from the spectrograms. As there exists a semantic gap between low-level and high-level features of the vanilla U-Net and its ability to handle only short-term dependencies, the audio-visual SE model in [80] incorporates RNN Attention (RA) blocks and the Res Paths (RPs) in the skip connections to reduce these gaps and increase the receptive field to enable it to find long-term dependencies. In conventional AV models speech and video are processed by separate encoders and then concatenated together before entering the decoder side. This not only increases the number of encoder parameters but also increases the burden on the decoder. In [80] early fusion of audio and video is done and a single encoder is used to avoid the above-mentioned problems. The proposed algorithm outperforms vanilla U-Net in terms of rejecting competitive speech and non-speech noises at much reduced computational cost.

The salient features of the source separation models discussed above are summarized in Table 3.

4) SOURCE INPAINTING

In dynamically changing noisy conditions, transient loud noise intrusions can lead to inescapable loss of information. Inspired by the image inpainting technique, where the damaged or missing parts of an artwork are filled in; in audio inpainting, the missing or severely degraded parts of spectrogram of audio are retrieved. In the past, the signal processing extrapolation schemes (e.g. [147]) were mostly used for recovering from such data losses in the time domain. In these methods, missing (i.e. the lost) samples are predicted by the past and future samples e.g. in [148], [149], and [150]. But these methods fail to recover the samples, distorted from random and irregular masks (intrusions) of arbitrary shapes [151].

a: SPEECH

The source inpainting model of [151], using U-Net and log-scaled complex STFT spectrograms, can recover time, frequency, or random gaps of up to 40% in a spectrogram of a one-second-long speech signal. SpeechVGG is used during

the training phase for estimating the training loss of U-Net, by comparing the predicted output spectrogram with the GT and feeding this estimate back to U-Net to update its training parameters. The model provides better speech quality and intelligibility, in case of extreme loss, when compared to the conventional inpainting methods. The model can recover parts of the spectrogram that are being missing and distorted by the additive noise or intrusions, that are as large as 400ms (i.e. the duration of a syllable or a short word) and 3.2 kHz in bandwidth. In the audio inpainting model of [92], long audio gaps > 200ms are filled, by using the pre-trained image network ResNet50 [152], after fine-tuning it with the masked STFT spectrograms. The output spectrograms of recovered speech are evaluated against the original spectrogram by using the objective evaluation methods commonly used for image and video inpainting algorithms such as the L1 loss and the perceptual loss. This was the first model to incorporate a pre-trained network for audio inpainting. The inpainting model of [153] is based on a well-known pix2pix image translation network [154] with a modified loss function, using U-Net on its generator side. The system reconstructs only the log-magnitude spectrogram while for phase reconstruction Griffin-Lim [155] algorithm is used. This network is shown to provide packet loss concealment of up to 120ms. The speech inpainting model in [156] is similar in architecture to [153] but it uses two strategies for phase reconstruction: 1) Phase Gradient Heap Integration (PGHI) [157] for the areas with high magnitudes (usually lower frequencies) and random phases for those with low magnitudes (mostly for high frequencies) since the higher frequencies do not contribute much to speech intelligibility. This reduces the buzz introduced if only PGHI were applied or the buzz caused by the [153] algorithm.

In the Internet of Things (IoT) era, wearable devices generally rely on environmental energy harvesting to alleviate the expensive maintenance overhead of battery recharging, but due to weak and unstable power supplies from these energy sources, these devices face intermittent failures. The model of [38] employs U-Net and STFT spectrograms to solve the intermittent speech loss problem, transmitted from such devices. First, the interpolation of missing segments is carried out in the time domain, followed by its spectrogram inpainting by U-Net. The results show tremendous improvement in quality, intelligibility, and Word Error Rate (WER) of the recovered speech, for devices that turn off intermittently for a duration as long as 128ms, after turning on only for 71ms.

b: MUSIC

The repetition of distinct patterns (themes, melodies, rhythms), makes the inpainting of long segments of music much easier than the inpainting of an A-periodic signal, e.g. speech [158]. The music inpainting model of [159], uses U-Net as the generator of the Generative Adversarial Network (GAN); another deep neural network, and splits the complex STFT spectrogram in two channels (real and imaginary) to be given at input of the model, which inpaints successfully the musical records, with pauses as long as 100ms. The inpainting model of [160] uses deep complex U-Net and complex-valued STFT spectrograms as input to estimate the complex-valued Ratio Mask (cRM), to restore the gaps due to hiss, clicks, thumps, and other common additive disturbances from old analog gramophone discs. The inpainting model proposed in [160] is also used in [161] for inpainting tape and cassette recordings. The models proposed in [162] and [163] are U-Net-based music inpainting models using CQT as their input. The model in [162] is only designed for piano sound for gaps up to 1.5s while that of [163] is for multi-instrument dataset for gaps up to 300ms. In both models, the time axis is downscaled by U Net while the frequency axis is not compressed on the encoder side. It is observed in [162] that for very long gaps, although the generated events are often reasonable, they do not align with the musical context so in [163], the CQT spectrogram is split into octave-wise sub transforms and they are processed one-by-one down the layers of U-Net. The instrument inpainting model in [81] is also a U-Net-based model using ResBlocks, taken from the ResNet. These blocks have local skip connections between convolutions on each level resulting in a network being capable of inpainting upto 1.6s duration of signal loss.

The salient features of the source inpainting models discussed above are summarized in Table 4.

5) SOURCE DENOISING AND DEREVERBERATION

In the source separation task, the noise source (unwanted signal) is focused, while denoising usually refers to the methods of removing the noise of diffuse characteristics e.g. noise in the market, café, random white noise, babble, wind, and airplane sounds. Most of the classical denoising techniques are based on statistical assumptions and so they fail to generalize well for the intrusive and non-stationary noise types. DNNs have made a breakthrough in this situation because of their ability to remove most of the background noise, regardless of its intensity and type [55].

Reverberation is also a kind of noise, produced by the source itself [1]. It is also diffuse in nature [173] and so acts as a diffuse noise source.

a: SPEECH

In the denoising model of [174], U-Net is trained on noisy squared log magnitude spectrograms, using the cleaned ones as GT, for single-channel speech enhancement. The U-Net architecture is the same as the generator architecture of pix2pix image GAN [154]. The denoising model of [93] again implements U-Net similar in architecture to the generator of pix2pix image GAN but with an ordinary STFT spectrogram. The model in [93] outperforms other deep learning denoising systems by offering better speech quality for AE applications and offers fewer errors in Automatic Speaker Verification (ASV) applications in the presence of white, café, airplane,

TABLE 4. Source inpainting models.

Model	Application	Pre-trained model	Network architecture	Input representation	Dataset used
			Speech inpainting		
M. Kegler et al. [151]	Speech inpainting	SpeechVGG	Both Speech VGG, and Vanilla	Log-scaled complex STFT spectrograms	LibriSpeech corpus [59]
Y. Chang et al. [92]	Speech and natural sounds inpainting	1). VGG16 2). ResNet50		STFT magnitude spectrogram	1). SC09 dataset of human voice [164] 2). ESC-50 dataset of natural sound [165]
C. Aironi et al. [153]	Speech inpainting	×	Vanilla U-Net in generator of pix2pix GAN [154]	Log magnitude STFT spectrogram	VCTK Corpus [166]
H. Zhao et al. [156]	Speech inpainting	×	Vanilla in generator of pix2pix GAN [154]	Log magnitude STFT spectrogram	LibriSpeech corpus [59]
Y.C. Lin et al. [38]	Speech inpainting	×	Deep complex U-Net	Complex STFT spectrograms	VCTK-DEMAND corpus [167]
			Music inpainting		
Y. Li et al. [159]	Music inpainting	×	Vanilla U-Net	Complex STFT spectrograms	Public Domain Project [168]
E. Moliner et al. [160]	Music inpainting	×	Vanilla U-Net	Complex STFT spectrograms	1.) The Great 78 Project [169] 2). MusicNet dataset [170] 3). Orchestral and opera recordings [171]
I. Irigaray et al. [161]	Tape and cassette recordings	×	Vanilla U-Net with intermediate layers at both sides consisting of DenseNet blocks	STFT spectrogram	1). MusicNet [170] for clean audio 2). Real recorded data for noise
E. Moliner et al. [162]	Piano inpainting	×	Vanilla U-Net with intermediate layers at both sides consisting of Residual Block (RBlock)	CQT	MAESTRO dataset [172]
E. Moliner et al. [163]	Music inpainting	×	Vanilla U-Net with intermediate layers at both sides consisting of Residual Block (RBlock)	CQT	MusicNet dataset [170]
K. Liu et al. [81]	Music inpainting	×	Vanilla U-Net with intermediate layers at both sides consisting of Residual Block (RBlock)	Mel Spectrogram	1). MusicNet dataset [170] 2). MAESTRO dataset [172]

babble, and market noise, under moderate SNRs, ranging from 5 to 15dB. In the model of [175], wind noise mixed with single-channel speech recorded outdoors is removed from its STFT spectrogram by U-Net. This wind noise subtraction model shows superior performance than the minimum statistics-based and nonnegative matrix factorization-based methods, under various SNR conditions. Conventional U-Net usually has a large number of trainable network parameters (ranging from 10 million to 100 million), which makes real-time execution on a typical smart device unfeasible [176] The denoising model proposed in [82], uses standard STFT and lightweight U-Net architecture, proposed by Google's Inception V4 networks [177], and achieves performance similar to conventional U-Net, while reducing the network's footprint size to 3.72% of the size of the conventional U-Net [82]. To utilize the phase information, the model in [97] uses similar architecture and complex STFT input representation for speech denoising, as used by the SVS

model [39]. The denoising model of [77] uses a spiking neural network in U-Net architecture. The individual neurons in this network emit a spike when their membrane potential reaches the threshold value. It is a low-powered network useful for cell phones. The noisy input magnitude spectrogram is mapped to a cleaner version by the network and later the noisy phases are combined to produce enhanced speech. The model in [178] uses the conventional U-Net architecture for speech enhancement and is found to be more effective under very low unseen SNR values and unseen noise types than the method using CNN.

In the dereverberation model of [1], U-Net is trained on the reverberated STFT spectrograms of monaural speech, using their clean counterpart as GT. However, the method fails to generate good results at longer reverberation times. The model of [179] is an online U-net structure for estimating the inverse filter response of each reflection path at each time unit, to better handle the time-varying reverberant conditions. This model is trained on Convolutive Transfer Function (CTF) coefficients arranged in a 2D matrix and provides better dereverberation performance at different levels of reverberation time, unseen type of room environment, and static and time-varying reverberant conditions for simulated and real rooms. The standard U-Net does not respond well to the train/test mismatch acoustic conditions. Implementing probabilistic bottleneck instead of deterministic in U-Net, in the denoising and dereverberation model of [65], enables it to adhere well to the unknown noise and reverberant scenarios, than the standard U-Net. The model is trained on log-scaled STFT power spectrograms.The use of symmetric filters in U-Net makes sense for regular images, as there is no difference between their x and y-axis. However, in the dereverberation model of [66], it is found that designing asymmetric filters, which have higher dimensions in the frequency domain than in the time domain combat the echoes better. The results on U-Net and GAN (using U-Net as its generator), with asymmetric filters and log-scaled STFT input representation, show reduced distortion in the output speech signal. The model presented in [180] is a binaural spatial cue-based dereverberation, using U-Net trained on spatial spectrograms. This model exploits the fact that there exists a distinction between the spatial cues of the direct path and the spatial cues of reverberations ([181] and [182]), so they can be separated. As the spatial cues generated by a source depend on its location, the moving sources or the sources placed at locations, other than those in the knowledge of a spatial cue-based dereverberation model, would not be dereverberated properly. This model uses beamforming at its front end, supported by U-Net at the backend, to learn the spatial cues of echoes and direct paths. After training, the beamformers are replaced with the binaurally separated microphones. This model has surpassed both the signal processing [183] and the RNN-based deep learning approaches [184], in terms of providing higher intelligibility and lesser distortion. As it is a spatial-cue-based algorithm, so the network is sensitive to the speaker's position. However, it is found that it is resilient to mild movements of the speaker up to 15cms in either x, y, or z direction, in the vicinity of its training position. Spectrum encoding as magnitude/ phase has shown better performance in anechoic conditions than its real/ imaginary representation which generalizes better in reverberant conditions. The complex U-Net model proposed in [185] is a denoising and dereverberation speech model that uses a variational latent space model and magnitude and phase spectrograms in the bottleneck of U-Net for dereverberation. The DNN models usually do not generalize well under unknown environments. This results in their poor performance under unknown conditions. The U-Net-based dereverberation and denoising model proposed in [67] is similar to conventional architecture except for two LSTM layers in the bottleneck park. It is a self-learning model that learns the spatial features of the environment from the input signal improving its adaptability under unknown environments. The deep learning models usually do not perform well when the recording conditions of the training and the test datasets do not match. Non-Matrix Factorization (NMF); a well-known machine learning technique, works well for such unseen conditions. In the speech-denoising model of [186], NMF is combined and jointly optimized with the U-Net model having a Temporal Activation layer (TAU-Net) to suppress temporal activations estimated by TAU-Net in unseen noisy conditions. The model outputs speech with better quality as compared to vanilla U-Net and SEGAN [41] in unseen conditions. To capture long-term dependencies, the U-Net-based SE model uses dilated convolutions to widen the receptive fields and maintain the TF resolution of feature maps at all levels of encoder and decoder. The system produces output with better quality and intelligibility than the LSTM and gated dilated convolutional networks.

The use of the generative diffusion model is a recent trend in natural image generation [187]. They are shown to perform better than GAN models, which capture less diversity and are difficult to train, collapsing without proper selection of hyper-parameters and regularizers, and difficult to scale and apply for unseen domains [187]. GANs tradeoff diversity for fidelity, producing high-quality samples but unable to cover the whole distribution. Diffusion models are a class of likelihood-based models that not only generate high-quality images but also offer other desirable characteristics such as distribution coverage, easy scalability, and a stationary training objective. For audio tasks these models are used for generating human-like natural language [188], highly diverse speech [189] and music [190], and voice conversion [191]. Whether used for image or audio, the core idea behind enhancement by diffusion model is to gradually convert the clean data to pure noise by gradually adding Gaussian noise to it in the forward process and then inverting the diffusion process by estimating the noise signal in the reverse process and using this estimated signal to restore the clean signal by subtracting it from the noisy data step by step [192]. The model of [193] has pioneered the use of diffusion network for AE and the models proposed in [193] and [194] are

based on direct processing of signals in 1D. The model of [195] has used the diffusion networks for the first time on spectrograms using deep Complex U-Net architecture and has outperformed the models of [193] and [194]. Later the models proposed in [192] and [196] have also used spectrograms as input. The model proposed in [192] uses the Deep Complex Convolution Recurrent Network (DCCRN) architecture of [197], while the model of [196] uses the Noise Conditional Score Network (NCSN++) of [198] based on multi-resolution U-Net architecture. It exceeds the model of [195] in performance. A complete review covering diffusion networks for speech enhancement and generation can be found in [23] so we will not cover these models any further.

b: OTHER AUDIO APPLICATIONS

Apart from speech and music, other audio applications e.g. anomaly detection by sound (whether in machine or auscultation), source localization, or environmental sound classifiers require audio denoising to achieve better accuracy. Now, we discuss a few models for these applications using U-Net for audio denoising.

c: SINGLE-TONE SOURCE LOCALIZATION

The multiple source localization model [199], uses 2D beamforming colored maps as input representation, and the corresponding 2D colored target maps as GTs, for training the U-Net. The model produces an Average Root Mean Square Error (RMSE) of just 2cm. This multi-source localization model requires neither any prior information about the number of sources to be localized, nor their presence necessarily on or near the predefined grid points in the coverage space of the beamformer. However, this model is only tested for sources generating a single tone.

d: HEALTH CARE

The use of highly sensitive and stable instrumentation, along with auscultation carried in optimal conditions, is recommended for reducing the noise, as its presence may result in incorrect classification of the pathology. But it is not always possible to obtain the ideal condition for the measurement and therefore noise is unavoidable, requiring denoising of the signal. The model in [200] tested both the Denoising CNN (DnCNN) [201] and U-Net for denoising the auscultation. The most interesting feature of this model is that the input layer of U-Net accepts a 2D signal directly in the time domain, without any transformation to the TF domain, by reshaping the audio vector to a 2D matrix, while the DnCNN is already designed to accept a 1D signal at its input. U-Net shows better denoising and requires less training time than the DnCNN. The model of [202] uses deep complex U-Net and complex STFT spectrograms for denoising of PhonoCardioGram (PCG) signal. The denoising model of [203], instead of being tested under the synthetic noisy conditions using Additive White Gaussian Noise (AWGN) noise, uses a more realistic approach and tests the model under four different noises including AWGN, additive pink Gaussian noise, speech, and real clinical background noise. This model improves the performance of automatic cardiac sound classification algorithms under very low SNR conditions. Speech has been discovered as a useful biomarker for the detection of COVID-19. The model proposed in [204] is used to denoise speech using conventional U-Net which is then classified for the absence or otherwise of COVID-19. The detection performance is improved by 10% especially under low SNR conditions as compared to the detection system without speech enhancement. The model proposed in [205] removes the clutter in ultrasound cardiac images caused by reverberations produced by sound reflections from echogenic structures such as subcutaneous fat, skin, bone, lung, cartilage, intercostal muscle, and out-of-scan-plane heart tissue. These reverberations appear as cloud-like diffuse haze and can affect the accuracy of diagnosis [205]. The model proposed in [205] uses U-Net and causal U-Net for haze removal. Causal U-Net is suitable for real-time inference while conventional U-Net is suitable for recorded data.

e: INDUSTRIAL SOUNDS

The ensemble model in [206] uses the log-scaled STFT spectrogram and U-Net in the pre-processing step of denoising the machine sounds. The U-Net is expected to reconstruct (inpaint) normal data even if few cues are available e.g. in noisy conditions. It generates an output mask, which is used for the retrieval of anomalous events. However, the system performance on the Detection and Classification of Acoustic Scenes and Events (DCASE) challenge 2022 industrial dataset [207] was poor, except for a few machines. Similar to [206], the noise removal step of the machine anomaly detection model, proposed in [208], uses a log-scaled complex Mel spectrogram, as an input to a deep complex U-Net. In [209], again U-Net is used not only in its conventional architecture but also in its two modified forms: 1) Nested U-Net [210] (where additional encoder-decoders are nested between the original encoder-decoders of the basic U-Net) and 2) ResU-Net (where residual connections exist in both the encoder and decoder side, which help in addressing the problem of gradient vanishing and thus produces more accurate results). These architectures provide much better performance in detecting the rare sound event in the audios of the industrial dataset of DCASE 2022 in case of extremely low SNRs than the systems without the denoising frontends. The models in [211] and [212] use U-Net and U-Net++ respectively for denoising the spectrograms of the sound from the planetary gearboxes which are widely used in many industrial categories such as mining, wind power generation, metal forming, etc. These TF spectrograms are later used for fault diagnoses. In [211] generalized S-transform is used for generating the TF map from the accelerometer recordings. Data was collected from the real environment, where the speed of gearboxes varied continuously. Varying speed

operations and a continuously changing environment are more challenging for fault diagnosis. U-Net has outperformed the classical methods in reducing the number of false positives and the processing time, making it a better choice for meeting the real-time requirements of fault diagnosis. In [212], an improved version of U-Net++ using Tversky loss [213] as an optimization objective is utilized for further improving the segmentation F1 score from 0.942 (in [211]) to 0.949 (in [212]).

f: ENVIRONMENTAL SOUNDS

In the case of environmental sounds, it is rare for any sound to be present alone. The model proposed in [214], first deploys an image CNN for sound event detection from the STFT spectrogram and then uses U-Net to predict the masks for segmenting different sources from the mixture's spectrogram. It can separate up to 75 different environmental sounds. However, this model works only for non-overlapping sounds present in the audio mixture.

To separate the background from the foreground sound of rare events, the sound event detection model of [68] uses two U-Net architectures: 1) the conventional U-Net, and 2) a novel U-Net architecture called U-Net with Limited Upsampling (LUU-Net) which applies limited upsampling on the decoder side to restore the original time axis and only a limited frequency axis. This saves the computational cost by 40% as compared to the conventional U-Net without any information loss as the onset and offset information of the rare event lies on the time axis. The segmentation masks obtained at the output of U-Net are used for obtaining the type and the timing information of the event from the weakly labeled dataset of DCASE 2018 tasks 1 [215] and 2 [216].

Another interesting of U-Net denoising is in underground utility tunnels where the condensation in summer may cause electric sparks in aged and corroded wiring which may result in fire. Installing CCTV cameras may not always help due to being expensive and due to the presence of blind spots so an acoustic-anomaly detection system is presented in [217] for detecting sparks. The spark sound is usually accompanied by the noise of ventilation fans inside the tunnel, the sound of falling water into the sump pit, and traffic noise over the road, thus difficult to be identified. Conventional U-Net is used for denoising the ambient sound which is later classified for the presence or otherwise of electric spark.

Supervised DNNs require both clean and noisy audio samples for training. However, the real audio recordings come with noises that cannot be separated to produce desired training samples. Secondly, most DNNs are trained on artificially created data using AWGN for noise that does not represent natural noise. The bird sound denoising model of [69] uses U-Net and U^2-Net [76] to separate the noise from the real recordings of the birds' sounds by transforming the audio into an STFT matrix. The sound files [76] have noises from the wind, waterfall, rain, and other natural sources. The GT images required for training are obtained by applying a threshold over the noisy areas of the STFT absolute spectrogram. The results show that U-Net performs better than U^2-Net.

The salient features of the source denoising and dereverberation models discussed above are summarized in Table 5.

IX. POTENTIAL DIRECTIONS FOR FUTURE RESEARCH

Although a lot has been explored, few areas still open for research using image U-Nets for AE can be summarized as below.

- As already explained above, different applications require different input representations. For example, as shown in Figure 5(a), the CQT representation is the most effective one in discriminating the two instruments in an audio mixture, as compared to others. Similarly, Mel spectrogram and scalogram are effective for detecting anomalies in open environment and health respectively (as shown in Figure 4). However, for large classes of bioacoustics and environmental sounds, the most appropriate representations are still missing [53]. Further investigation is required to find them for these sounds.
- Exploring new pretrained models from the already available large repository of networks trained for computer vision for the task of MSS, music inpainting, source dereverberation, and denoising where no such significant examples of using them exist.
- Till now, for the AE tasks requiring spectrogram segmentation, the spectrogram given at the input of U-Net is merely a 2D matrix of numbers. This can be regarded as a grayscale image. Although colored spectrograms have been tried for the AE classification tasks (e.g. [240] and [241]), to the best of our knowledge, there is no such example we have found for spectrogram segmentation applications like audio denoising, separation, or inpainting. The effect of changing the color map of the spectrogram on human perception is evident [48], and has already proved impactful in the classification of lung sounds [240] and in automatic speech emotion recognition (SER) [241]. It is also visible from Figure 5(b) that 'parula' and 'colorcube' color maps of the CQT spectrogram show the contents of each instrument better than the 'prism' color map, which fails to provide any discrimination of two sources whatsoever in the region of mixing. Similarly, in Figure 6, the effect of using different color maps for the speech denoising application is shown, where even by visual inspection of clean and noisy spectrograms, the noise is easily differentiable from the regions of clean speech in 'parula' color map, mildly in 'pink', and poorly in the 'prism' color-map. Now the question arises, does working with colors instead of just using a 2D TF matrix make any difference in the performance of

TABLE 5. Source denoising and dereverberation models.

Model	Application	Pre-trained model	Network architecture	Input representation	Dataset used
Speech denoising and dereverberation					
A. E. Bulut et al. [174]	Speech denoising	×	Generator of pix2pix [155]	Squared log magnitude spectrograms	1). Voice Bank corpus [144] 2). DEMAND [145]
D. Michelsanti et al. [93]	Speech denoising	×	Vanilla U-Net with intermediate layers at both sides consisting of DenseNet blocks	STFT magnitude spectrogram	1). TIMIT [131] 2). RSR2015 [218]
G.W.Lee et al. [175]	Speech denoising	×	Vanilla U-Net	FFT arranged in 2D matrix	1). From Internet 2). Recordings in soccer field and parking area
K. M. Jeon et al. [82]	Speech denoising	×	Vanilla U-Net with intermediate layers at both sides consisting of Multi-Lane Dimensionality Reduction (MLDR) module	STFT magnitude	1). Voice Bank corpus [144] 2). DEMAND [145]
H. S.Choi et al. [97]	Speech denoising	×	Deep complex U-Net	Complex STFT spectrograms	1). Voice Bank corpus [144] 2). DEMAND [145]
A. Riahi et al. [77]	Speech denoising	×	U-Net with Leaky Integrate-and-Fire (LIF) neuron model	Squared log magnitude spectrograms	1). Voice Bank corpus [144] 2). DEMAND [145]
K. Akter et al. [178]	Speech denoising	×	Vanilla U-Net with intermediate layers at both sides consisting of attention mechanism	STFT magnitude	TIMIT corpus [131]
K. Nakazawa et al. [1]	Speech dereverberation	×	Vanilla U-Net	Squared log magnitude spectrograms	1). Real speech recordings 2). Simulated Room Impulse Responses (RIRs) [219]
H Chung et al. [179]	Speech dereverberation	×	Vanilla U-Net	Convolutive Transfer Function (CTF) coefficients arranged in 2D matrix	1). TIMIT corpus [131] 2). Simulated RIRs [220] 3). Real RIRs from the C4DM database [221].
E. J. Nustede et al. [65]	Speech denoising + dereverberation	×	Vanilla U-Net with dilated convolutions with variational bottleneck	Log-scaled power spectrogram	Deep Noise Suppression Challenge 2020 [222]
O. Ernst et al. [66]	Speech denoising + dereverberation	×	Both U-Net with asymmetric filters, and U-Net in generator of GAN	Log-scaled magnitude spectrograms	REVERB challenge [223]
S. Gul et al. [180]	Speech denoising + dereverberation	×	Vanilla U-Net	Spatial spectrograms	1). TIMIT [131] 2). Real RIRs from the University of Surrey [224].
E. J. Nustede et al. [185]	Speech denoising + dereverberation	×	Complex U-Net with dilated convolutions and variational bottleneck	Complex STFT spectrogram	1). Deep Noise Suppression Challenge 2020 [222] 2). Voice Bank corpus [144] 3). DEMAND [145]
S. Gao et al. [67]	Speech denoising + dereverberation	×	Vanilla U-Net with BLSTM bottleneck	Complex STFT spectrogram	LibriSpeech corpus [59]
K. M. Jeon et al. [186]	Speech denoising	×	Vanilla U-Net with BLSTM bottleneck	STFT magnitude	1). Voice Bank corpus [144] 2). DEMAND [145]
T. Grzywalski et al. [225]	Speech denoising	×	Vanilla U-Net	STFT magnitude	1). WSJ0 [226] 2). TIMIT [131] 3). Freesound [227] 4). Noisex [228]

TABLE 5. *(Continued.)* **Source denoising and dereverberation models.**

S. Welker et al. [195]	Speech denoising + dereverberation	×	Deep complex U-Net	Complex STFT spectrogram	1). Voice Bank corpus [144] 2). DEMAND [145]
J. Richter et al. [196]	Speech denoising + dereverberation	×	NCSN++ [198]		1). Voice Bank corpus [144] 2). DEMAND [145] 3). WSJ-0 [226]
Other audio applications					
Lee et al. [199]	Source localization	×	Vanilla U-Net	Beamforming map	Simulation
T. S. Sharan et al. [200]	Denoising auscultation sounds	×	Vanilla U-Net	1D audio vector shaped to 160*256	PhysioNet challenge 2017 [229]
A. Mukherjee et al. [202]	Denoising auscultation sounds	×	Vanilla U-Net	Complex STFT spectrogram	PASCAL heart sound recordings [230]
C. González et al. [203]	Denoising auscultation sounds	×	Vanilla U-Net	STFT magnitude spectrogram	1). [231] 2). PhysioNet challenge 2017 [229] 3). LibriSpeech [59]
S. Liu et al. [204]	Speech denoising for COVID 19 detection	×	Vanilla U-Net	STFT spectrogram	1). AudioSet [61] 2). DiCOVA [232]
T. S. Jahren et al. [205]	Clutter removal in cardiac ultrasound	×	Vanilla U-Net with intermediate layers at both sides consisting of Residual Block (RBlock)	Log-scaled images	Real recordings
J. Yamashita et al. [206]	Audio denoising inpainting	×	Vanilla U-Net	Log-scaled magnitude spectrograms	DCASE 2022 Challenge Task 2 [207]
P. Daniluk et al. [208]	Audio denoising	×	Deep Complex U-Net	Log-scaled complex Mel spectrograms	DCASE 2020 Challenge Task 2 [233]
Y. Shin et al. [209]	Audio denoising	×	Vanilla U-Net	Mel spectrograms	DCASE 2022 Challenge [207]
P. Zhang et al. [211]	Fault diagnosis in planetary gearboxes	×	Both Vanilla U-Net , and U-Net++ [210]	S-Transform	Real machine recordings
P. Zhang et al. [212]	Fault diagnosis in planetary gearboxes	×	Improved U-Net++	S-Transform	Real machine recordings
Y. Sudo et al. [214]	Environmental sound classification and segmentation	×	Vanilla U-Net	STFT spectrogram	1). ATRECSS — ATR English Speech Corpus For Speech Synthesis [234] 2). RWCP Sound Scene Database [235] 3). Freesound General-Purpose Audio Tagging Challenge [236] 4). DCASE 2016 Task 2 [237] 5). RWC-music database [238] 6). Freesound [227]
S. Lee et al. [68]	Acoustic scene classification and rare event detection	×	U-Net architecture with Limited Upsampling (LUU-Net)	STFT spectrogram	DCASE 2018 Challenge Task 1 [215], Task2 [216]
B.-J. Lee et al. [217]	abnormal sound detection in underground utility tunnels	×	Vanilla U-Net	STFT spectrogram	Artificially created
Y. Zhang et al. [69]	Birds sound denoising and speech and audio denoising and noise estimation	×	U^2-Net	Absolute STFT spectrogram	Xeno-canto [239]

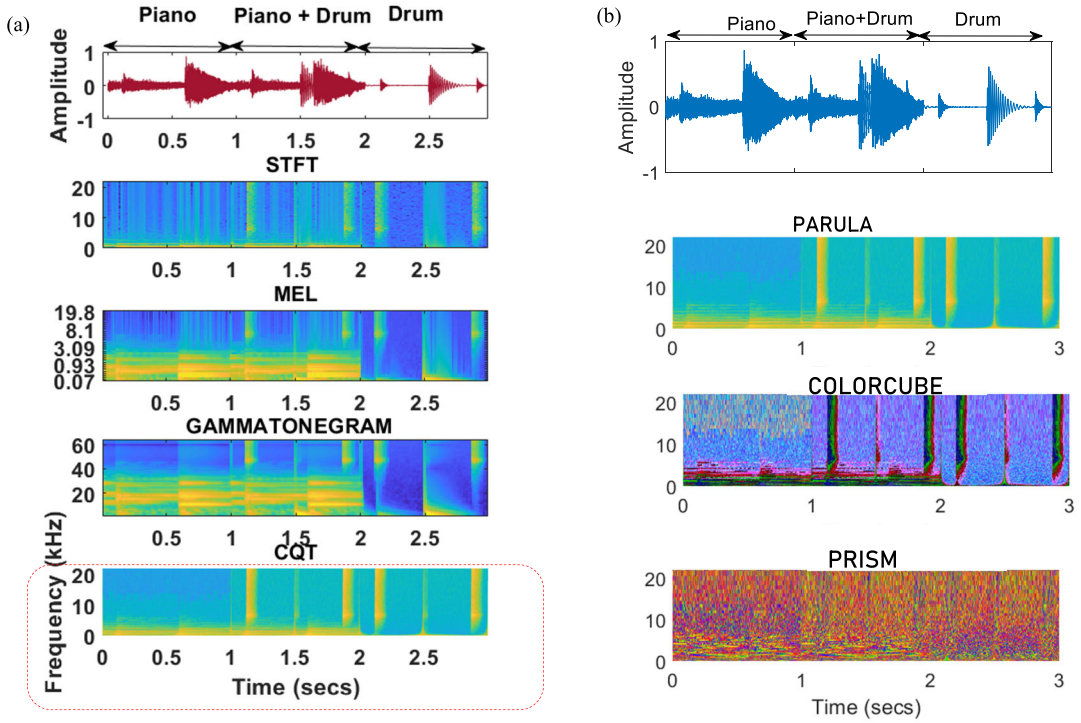

FIGURE 5. Effect of changing (a) the type, and (b) the color-map of CQT spectrogram on separation of instruments

audio enhancement networks? What if the spectrogram is given as a colored image with 3 dimensions using either Red Green Blue (RGB), YCbCr (Y′ is the luma component and Cb and Cr are the blue-difference and red-difference chroma components), or Hue Saturation Value (HSV) color space? Does it make any impact on the quality of enhanced audio by increasing each pixel's information from a single to a three-digit (one for each of R, G, and B, or Y, Cb, and Cr or H, S, and V) tuple? Different color spaces have proved their worth in different applications. For example, in varying illumination conditions and complicated backgrounds, YCrCb is beneficial for face [242] and jaundice detection [243], HSV for detection of broken stitches in industrial sewing machines [244], fire detection [245], and RGB for detection of cancer cells from biopsy images [246] and anemia [247]. Although, by using colored spectrograms as input to U-Net the network parameters would certainly increase but will it reduce the learning epochs of networks? These questions open a new direction for future research, relying heavily upon the extensive research and tools available in the already-established field of processing colored images. Our initial experiments with colors for speech denoising show them to be highly effective in reducing the computational cost and time of the system's training without any depreciation in the model's output speech quality when compared with state-of-the-art systems [248].

- Using lightweight U-Net architectures for AE in mobile applications. Examples of U-Net architectures with fewer parameters useful for image segmentation are 1) Efficient and Lightweight U-Net (ELU-Net) (developed for medical image segmentation [249]), 2) Attention U-Net and SqueezeNet (ATT Squeeze) U-Net (developed for forest fire detection [250]), 3) Lighter U-net @128 (developed for lesion segmentation in ultrasound images in [251]), and 4) lightweight U-Net (developed for detection and segmentation of iron ore green pellets in [252]).

X. CONCLUSION

In this article, we have presented a review, focused entirely on the use of U-Nets for the AE applications. Although the conversion of audio to time-frequency domain and its advantages are already well established, treating these 2D representations as images, and utilizing U-Net, is an approach, that has outperformed many of the state-of-the-art signal processing and machine learning algorithms for AE applications. The use of U-Nets for AE is currently in its childhood, yet it is enjoying explosive interest from researchers. In the future, investigating the effect of different color maps and color spaces of input representations, merging the 2D deep learning models with other machine learning algorithms, and exploring the most optimal input representation and more lightweight variants and pretrained models of U-Net for mobile devices will benefit more AE applications.

ACKNOWLEDGMENT

Open Access funding for this article was provided by the Qatar National Library.

REFERENCES

[1] K. Nakazawa and K. Kondo, "De-reverberation using CNN for non-reference reverberant speech intelligibility estimation," in *Proc. Int. Congr. Acoust.*, Aachen, Germany, Sep. 2019, pp. 3098–3102.

[2] J. Turian, J. Shier, G. Tzanetakis, K. McNally, and M. Henry, "One billion audio sounds from GPU-enabled modular synthesis," in *Proc. 24th Int. Conf. Digit. Audio Effects (DAFx)*, Sep. 2021, pp. 222–229.

[3] M. Cartwright and B. Pardo, "SynthAssist: An audio synthesizer programmed with vocal imitation," in *Proc. 22nd ACM Int. Conf. Multimedia*, Nov. 2014, pp. 741–742.

[4] D. Michelsanti, Z.-H. Tan, S.-X. Zhang, Y. Xu, M. Yu, D. Yu, and J. Jensen, "An overview of deep-learning-based audio-visual speech enhancement and separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 1368–1396, 2021.

[5] C. C. Tappert, "Who is the father of deep learning?" in *Proc. Int. Conf. Comput. Sci. Comput. Intell. (CSCI)*, Las Vegas, NV, USA, Dec. 2019, pp. 343–348.

[6] H. Purwins, B. Li, T. Virtanen, J. Schlüter, S.-Y. Chang, and T. Sainath, "Deep learning for audio signal processing," *IEEE J. Sel. Topics Signal Process.*, vol. 13, no. 2, pp. 206–219, May 2019.

[7] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.

[8] Y. LeCun, K. Kavukcuoglu, and C. Farabet, "Convolutional networks and applications in vision," in *Proc. IEEE Int. Symp. Circuits Syst.*, May 2010, pp. 253–256.

[9] K. Fukushima and S. Miyake, "Neocognitron: A new algorithm for pattern recognition tolerant of deformations and shifts in position," *Pattern Recognit.*, vol. 15, no. 6, pp. 455–469, Jan. 1982.

[10] Y. Le Cun et al., "Handwritten digit recognition with a back-propagation network," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 2, 1989, pp. 1–9.

[11] P. Sharma and A. Singh, "Era of deep neural networks: A review," in *Proc. 8th Int. Conf. Comput., Commun. Netw. Technol. (ICCCNT)*, Jul. 2017, pp. 1–5.

[12] D. Tang, B. Qin, and T. Liu, "Document modeling with gated recurrent neural network for sentiment classification," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Sep. 2015, pp. 1422–1432.

[13] K. Choi, G. Fazekas, M. Sandler, and K. Cho, "Convolutional recurrent neural networks for music classification," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 2392–2396.

[14] Z. Zuo, B. Shuai, G. Wang, X. Liu, X. Wang, B. Wang, and Y. Chen, "Convolutional recurrent neural networks: Learning spatial dependencies for image representation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2015, pp. 18–26.

[15] S. Sigtia, E. Benetos, and S. Dixon, "An end-to-end neural network for polyphonic piano music transcription," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 5, pp. 927–939, May 2016.

[16] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. 18th Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, Munich, Germany, Oct. 2015, pp. 234–241.

[17] D. Stoller, S. Ewert, and S. Dixon, "Wave-U-Net: A multi-scale neural network for end-to-end audio source separation," 2018, *arXiv:1806.03185*.

[18] R. Giri, U. Isik, and A. Krishnaswamy, "Attention Wave-U-Net for speech enhancement," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust. (WASPAA)*, Oct. 2019, pp. 249–253.

[19] H.-S. Choi, S. Park, J. H. Lee, H. Heo, D. Jeon, and K. Lee, "Real-time denoising and dereverberation with tiny recurrent U-Net," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Ontario, ON, Canada, Jun. 2021, pp. 5789–5793.

[20] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 10, pp. 1702–1726, Oct. 2018.

[21] S. Latif, H. Cuayáhuitl, F. Pervez, F. Shamshad, H. S. Ali, and E. Cambria, "A survey on deep reinforcement learning for audio-based applications," *Artif. Intell. Rev.*, vol. 56, no. 3, pp. 2193–2240, Mar. 2023.

[22] D. Hepsiba, R. Vinotha, and L. D. V. Anand, "Speech enhancement and recognition using deep learning algorithms: A review," in *Proc. Comput. Vis. Bio-Inspired Comput.*, Apr. 2023, pp. 259–268.

[23] C. Zhang, C. Zhang, S. Zheng, M. Zhang, M. Qamar, Su.-H. Bae, and I. S. Kweon, "A survey on audio diffusion models: Text to speech synthesis and enhancement in generative AI," 2023, *arXiv:2303.13336*.

[24] A. Mehrish, N. Majumder, R. Bharadwaj, R. Mihalcea, and S. Poria, "A review of deep learning techniques for speech processing," *Inf. Fusion*, vol. 99, Nov. 2023, Art. no. 101869.

[25] A. Golmakani, M. Sadeghi, and R. Serizel, "Audio-visual speech enhancement with a deep Kalman filter generative model," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2023, pp. 1–5.

[26] K. Zmolikova, M. Delcroix, T. Ochiai, K. Kinoshita, J. Cernocký, and D. Yu, "Neural target speech extraction: An overview," *IEEE Signal Process. Mag.*, vol. 40, no. 3, pp. 8–29, May 2023.

[27] R. Guo, Z. Luo, and M. Li, "A survey of optimization methods for independent vector analysis in audio source separation," *Sensors*, vol. 23, no. 1, p. 493, Jan. 2023.

[28] R. V. Devi and Vasundhara, "Review on recent advances in hearing aids: A signal processing perspective," in *Proc. Int. Conf. Paradigms Comput., Commun. Data Sci., Algorithms Intell. Syst.* Singapore: Springer, Feb. 2022, pp. 225–240.

[29] D. Mukhutdinov, A. Alex, A. Cavallaro, and L. Wang, "Deep learning models for single-channel speech enhancement on drones," *IEEE Access*, vol. 11, pp. 22993–23007, 2023.

[30] J. Zhang, J. Tang, and L.-R. Dai, "RNN-BLSTM based multi-pitch estimation," in *Proc. Interspeech*, 2016, pp. 1785–1789.

[31] A. Shewalkar, D. Nyavanandi, and S. A. Ludwig, "Performance evaluation of deep neural networks applied to speech recognition: RNN, LSTM and GRU," *J. Artif. Intell. Soft Comput. Res.*, vol. 9, no. 4, pp. 235–245, Oct. 2019.

[32] H. R. Guimarães, H. Nagano, and D. W. Silva, "Monaural speech enhancement through deep Wave-U-Net," *Expert Syst. Appl.*, vol. 158, Nov. 2020, Art. no. 113582.

[33] *2010—2019: The Rise of Deep Learning*. Accessed: Dec. 19, 2023. [Online] Available: https://thenextweb.com/news/2010-2019-the-rise-of-deep-learning

[34] *ImageNet*. Accessed: Dec. 20, 2023. [Online] Available: https://www.image-net.org/

[35] C. Macartney and T. Weyde, "Improved speech enhancement with the Wave-U-Net," 2018, *arXiv:1811.11307*.

[36] A. Bosca, A. Guérin, L. Perotin, and S. Kitic, "Dilated U-Net based approach for multichannel speech enhancement from first-order ambisonics recordings," in *Proc. 28th Eur. Signal Process. Conf. (EUSIPCO)*, Amsterdam, The Netherlands, Jan. 2021, pp. 216–220.

[37] A. Jansson, E. Humphrey, N. Montecchio, R. Bittner, A. Kumar, and T. Weyde, "Singing voice separation with deep U-Net convolutional networks," in *Proc. 18th Int. Soc. Music Inf. Retr. Conf.*, Suzhou, China, Oct. 2017, pp. 23–27.

[38] Y.-C. Lin, T.-A. Hsieh, K.-H. Hung, C. Yu, H. Garudadri, Y. Tsao, and T.-W. Kuo, "Intermittent speech recovery," 2021, *arXiv:2106.05229*.

[39] W. Choi, M. Kim, J. Chung, D. Lee, and S. Jung, "Investigating U-Nets with various intermediate blocks for spectrogram-based singing voice separation," 2020, *arXiv:1912.02591*.

[40] G. Dahy, M. A. A. Refaey, R. Alkhoribi, and M. Shoman, "A speech separation system in video sequence using dilated inception network and U-Net," *Egyptian Informat. J.*, vol. 23, no. 4, pp. 121–131, Dec. 2022.

[41] S. Pascual, A. Bonafonte, and J. Serra, "SEGAN: Speech enhancement generative adversarial network," 2017, *arXiv:1703.09452*.

[42] Y. Luo and N. Mesgarani, "TaSNet: Time-domain audio separation network for real-time, single-channel speech separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Calgary, AB, Canada, Apr. 2018, pp. 696–700.

[43] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 8, pp. 1256–1266, Aug. 2019.

[44] A. Défossez, G. Synnaeve, and Y. Adi, "Real time speech enhancement in the waveform domain," in *Proc. Interspeech*, Oct. 2020, pp. 3291–3295.

[45] J. C. Steinberg and N. R. French, "The portrayal of visible speech," *J. Acoust. Soc. Amer.*, vol. 18, no. 1, pp. 4–18, Jul. 1946.

[46] A. S. Bregman, *Auditory Scene Analysis*. Cambridge, MA, USA: MIT Press, 1990.

[47] I. Mcloughlin, *Applied Speech and Audio Processing*. Cambridge, U.K.: Cambridge Univ. Press, 2009.

[48] D. Ciric, Z. Peric, J. Nikolic, and N. Vucic, "Audio signal mapping into spectrogram-based images for deep learning applications," in *Proc. 20th Int. Symp. INFOTEH-JAHORINA (INFOTEH)*, East Sarajevo, Bosnia and Herzegovina, Mar. 2021, pp. 1–6.

[49] E. J. Humphrey and J. P. Bello, "Rethinking automatic chord recognition with convolutional neural networks," in *Proc. 11th Int. Conf. Mach. Learn. Appl.*, Dec. 2012, pp. 357–362.

[50] V. S. Kadandale, J. F. Montesinos, G. Haro, and E. Gómez, "Multi-channel U-Net for music source separation," in *Proc. IEEE 22nd Int. Workshop Multimedia Signal Process. (MMSP)*, Sep. 2020, pp. 1–6.

[51] L. Wang, H. Ding, and F. Yin, "Combining superdirective beamforming and frequency-domain blind source separation for highly reverberant signals," *EURASIP J. Audio, Speech, Music Process.*, vol. 2010, pp. 1–13, Dec. 2010.

[52] S. Rickard, "The DUET blind source separation algorithm," in *Blind Speech Separation* (Signals and Communication Technology). Dordrecht, The Netherlands: Springer, 2007.

[53] P. C. Bermant, "BioCPPNet: Automatic bioacoustic source separation with deep neural networks," *Sci. Rep.*, vol. 11, no. 1, p. 23502, Dec. 2021.

[54] E. Cano, D. FitzGerald, A. Liutkus, M. D. Plumbley, and F.-R. Stöter, "Musical source separation: An introduction," *IEEE Signal Process. Mag.*, vol. 36, no. 1, pp. 31–40, Jan. 2019.

[55] S. A. Nossier, J. Wall, M. Moniri, C. Glackin, and N. Cannings, "An experimental analysis of deep learning architectures for supervised speech enhancement," *Electronics*, vol. 10, no. 1, p. 17, Dec. 2020.

[56] S. Albawi, T. A. Mohammed, and S. Al-Zawi, "Understanding of a convolutional neural network," in *Proc. Int. Conf. Eng. Technol. (ICET)*, Antalya, Turkey, Aug. 2017, pp. 1–6.

[57] F. Chollet, *Deep Learning with Python*, vol. 6. Shelter Island, NY, USA: Manning, 2017.

[58] G. Gwardys and D. Grzywczak, "Deep image features in music information retrieval," *Int. J. Electron. Telecommun.*, vol. 60, no. 4, pp. 321–326, Dec. 2014.

[59] S. Nisar, O. U. Khan, and M. Tariq, "An efficient adaptive window size selection method for improving spectrogram visualization," *Comput. Intell. Neurosci.*, vol. 2016, Aug. 2016, Art. no. 6172453.

[60] J. Shor, A. Jansen, R. Maor, O. Lang, O. Tuval, F. D. C. Quitry, M. Tagliasacchi, I. Shavitt, D. Emanuel, and Y. Haviv, "Towards learning a universal non-semantic representation of speech," in *Proc. Interspeech*, Oct. 2020, pp. 140–144.

[61] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, New Orleans, LA, USA, Mar. 2017, pp. 776–780.

[62] D. Niizumi, D. Takeuchi, Y. Ohishi, N. Harada, and K. Kashino, "BYOL for audio: Self-supervised learning for general-purpose audio representation," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2021, pp. 1–8.

[63] E. Fonseca, X. Favory, J. Pons, F. Font, and X. Serra, "FSD50K: An open dataset of human-labeled sound events," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 30, pp. 829–852, Dec. 2022.

[64] S. Gul, M. S. Fulaly, M. S. Khan, and S. W. Shah, "Clustering of spatial cues by semantic segmentation for anechoic binaural source separation," *Appl. Acoust.*, vol. 171, Jan. 2021, Art. no. 107566.

[65] E. J. Nustede and J. Anemüller, "Towards speech enhancement using a variational U-Net architecture," in *Proc. 29th Eur. Signal Process. Conf. (EUSIPCO)*, Dublin, Ireland, Aug. 2021, pp. 481–485.

[66] O. Ernst, S. E. Chazan, S. Gannot, and J. Goldberger, "Speech dereverberation using fully convolutional networks," in *Proc. 26th Eur. Signal Process. Conf. (EUSIPCO)*, Rome, Italy, Sep. 2018, pp. 390–394.

[67] S. Gao, X. Wu, and T. Qu, "A physical model-based self-supervised learning method for signal enhancement under reverberant environment," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 31, pp. 2100–2110, 2023.

[68] S. Lee, H. Kim, and G.-J. Jang, "Weakly supervised U-Net with limited upsampling for sound event detection," *Appl. Sci.*, vol. 13, no. 11, p. 6822, Jun. 2023.

[69] 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.

[70] E. Perez, F. Strub, H. De Vries, V. Dumoulin, and A. Courville, "Film: Visual reasoning with a general conditioning layer," in *Proc. AAAI Conf. Artif. Intell.*, Apr. 2018, vol. 32, no. 1, pp. 3942–3951.

[71] G. Meseguer-Brocal and G. Peeters, "Conditioned-U-Net: Introducing a control mechanism in the U-Net for multiple source separations," in *Proc. Int. Soc. Music Inf. Retr. Conf. (ISMIR)*, Delft, The Netherlands, Nov. 2019, pp. 159–165.

[72] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2261–2269.

[73] X. Min, G. Zhai, J. Zhou, X.-P. Zhang, X. Yang, and X. Guan, "A multimodal saliency model for videos with high audio-visual correspondence," *IEEE Trans. Image Process.*, vol. 29, pp. 3805–3819, 2020.

[74] Y. Zhang, K. Wu, and M. Zhao, "An audio-visual separation model integrating dual-channel attention mechanism," *IEEE Access*, vol. 11, pp. 63069–63080, 2023.

[75] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: A nested U-Net architecture for medical image segmentation," in *Proc. 4th Int. Workshop Deep Learn. Med. Image Anal. Multimodal Learn. Clin. Decis. Support (DLMIA)*, Granada, Spain, Sep. 2018, pp. 3–11.

[76] X. Qin, Z. Zhang, C. Huang, M. Dehghan, O. R. Zaiane, and M. Jagersand, "U^2-Net: Going deeper with nested U-structure for salient object detection," *Pattern Recognit.*, vol. 106, Oct. 2020, Art. no. 107404.

[77] A. Riahi and É. Plourde, "Single channel speech enhancement using U-Net spiking neural networks," 2023, *arXiv:2307.14464*.

[78] C. Gan, D. Huang, H. Zhao, J. B. Tenenbaum, and A. Torralba, "Music gesture for visual sound separation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10475–10484.

[79] C. Huang, S. Liang, Y. Tian, A. Kumar, and C. Xu, "DAVIS: High-quality audio-visual separation with generative diffusion models," 2023, *arXiv:2308.00122*.

[80] J.-W. Hwang, R.-H. Park, and H.-M. Park, "Efficient audio-visual speech enhancement using deep U-Net with early fusion of audio and video information and RNN attention blocks," *IEEE Access*, vol. 9, pp. 137584–137598, 2021.

[81] K. Liu, W. Gan, and C. Yuan, "MAID: A conditional diffusion model for long music audio inpainting," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Rhodes Island, Greece, Jun. 2023, pp. 1–5.

[82] K. M. Jeon, C. Chun, G. Kim, C. Leem, B. Kim, and W. Choi, "Lightweight U-Net based monaural speech source separation for edge computing device," in *Proc. IEEE Int. Conf. Consum. Electron. (ICCE)*, Jan. 2020, pp. 1–4.

[83] J. Walsh, A. Othmani, M. Jain, and S. Dev, "Using U-Net network for efficient brain tumor segmentation in MRI images," *Healthcare Anal.*, vol. 2, Nov. 2022, Art. no. 100098.

[84] A. Lukin and J. G. Todd, "Adaptive time-frequency resolution for analysis and processing of audio," in *Proc. 120th Convention Paper*, Paris, France, May 2006, pp. 1–10.

[85] *How to Convert A—Mel Spectrogram to Log-Scaled Melspectrogram*. Accessed: Dec. 20, 2023. [Online] Available: https://datascience.stackexchange.com/questions/27634/how-to-convert-a-mel-spectrogram-to-log-scaled-mel-spectrogram

[86] C. De la Fuente, E. Martinez-Valdes, J. I. Priego-Quesada, A. Weinstein, O. Valencia, M. R. Kunzler, J. Alvarez-Ruf, and F. P. Carpes, "Understanding the effect of window length and overlap for assessing sEMG in dynamic fatiguing contractions: A non-linear dimensionality reduction and clustering," *J. Biomech.*, vol. 125, Aug. 2021, Art. no. 110598.

[87] A. J. R. Simpson, "Time-frequency trade-offs for audio source separation with binary masks," 2015, *arXiv:1504.07372*.

[88] J.-Y. Lee, "Variable short-time Fourier transform for vibration signals with transients," *J. Vibrat. Control*, vol. 21, no. 7, pp. 1383–1397, May 2015.

[89] *Time-Frequency Gallery*. Accessed: Dec. 20, 2023. [Online] Available: https://www.mathworks.com/help/signal/ug/time-frequency-gallery.html

[90] Y. Tagawa, R. Maskeliūnas, and R. Damaševičius, "Acoustic anomaly detection of mechanical failures in noisy real-life factory environments," *Electronics*, vol. 10, no. 19, p. 2329, Sep. 2021.

[91] *Scalogram Computation in Signal Analyzer*. Accessed: Dec. 20, 2023. [Online] Available: https://www.mathworks.com/help/signal/ug/scalogram-computation-in-signal-analyzer.html

[92] Y.-L. Chang, K.-Y. Lee, P.-Y. Wu, H.-Y. Lee, and W. Hsu, "Deep long audio inpainting," 2019, *arXiv:1911.06476*.

[93] D. Michelsanti and Z.-H. Tan, "Conditional generative adversarial networks for speech enhancement and noise-robust speaker verification," in *Proc. Interspeech*, Stockholm, Sweden, Aug. 2017, pp. 2008–2012.

[94] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 14, no. 4, pp. 1462–1469, Jul. 2006.

[95] S. A. Nossier, J. Wall, M. Moniri, C. Glackin, and N. Cannings, "Two-stage deep learning approach for speech enhancement and reconstruction in the frequency and time domains," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Padua, Italy, Jul. 2022, pp. 1–10.
[96] A. Cohen-Hadria, A. Roebel, and G. Peeters, "Improving singing voice separation using deep U-Net and Wave-U-Net with data augmentation," in *Proc. 27th Eur. Signal Process. Conf. (EUSIPCO)*, Coruna, Spain, Sep. 2019, pp. 1–5.
[97] H.-S. Choi, J.-H. Kim, J. Huh, A. Kim, J.-W. Ha, and K. Lee, "Phase-aware speech enhancement with deep complex U-Net," in *Proc. Int. Conf. Learn. Represent.*, Vancouver, BC, Canada, May 2018.
[98] E. V. George and V. P. Devassia, "A novel U-Net with dense block for drum signal separation from polyphonic music signal mixture," *Signal, Image Video Process.*, vol. 17, no. 3, pp. 627–633, Apr. 2023.
[99] H. Zhao et al., "The sound of pixels," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 570–586.
[100] H. Zhao, C. Gan, W.-C. Ma, and A. Torralba, "The sound of motions," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1735–1744.
[101] R. Gao and K. Grauman, "Co-separating sounds of visual objects," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3878–3887.
[102] S. Mo and P. Morgado, "A unified audio-visual learning framework for localization, separation, and recognition," 2023, *arXiv:2305.19458*.
[103] Z. Rafii, A. Liutkus, F.-R. Stöter, S. I. Mimilakis, and R. Bittner. (Dec. 2017). *MUSDB18—A Corpus for Music Source Separation*. Accessed: Dec. 20, 2023. [Online]. Available: https://hal.inria.fr/hal-02190845
[104] *MUSIC*. [Online]. Available: http://sound-of-pixels.csail.mit.edu/
[105] B. Li, X. Liu, K. Dinesh, Z. Duan, and G. Sharma, "Creating a multitrack classical music performance dataset for multimodal music analysis: Challenges, insights, and applications," *IEEE Trans. Multimedia*, vol. 21, no. 2, pp. 522–535, Feb. 2019.
[106] E. C. Cherry, *On Human Communication*. Cambridge, MA, USA: MIT Press, 1957.
[107] C. Bergler, H. Schröter, R. X. Cheng, V. Barth, M. Weber, E. Nöth, H. Hofer, and A. Maier, "ORCA-SPOT: An automatic killer whale sound detection toolkit using deep learning," *Sci. Rep.*, vol. 9, no. 1, Jul. 2019, Art. no. 10997.
[108] S. Gul, M. S. Khan, and S. W. Shah, "Integration of deep learning with expectation maximization for spatial cue-based speech separation in reverberant conditions," *Appl. Acoust.*, vol. 179, Aug. 2021, Art. no. 108048.
[109] S. Gul, M. S. Khan, N. B. Yoma, S. W. Shah, and Sheheryar, "Enhancing the correlation between the quality and intelligibility objective metrics with the subjective scores by shallow feed forward neural network for time–frequency masking speech separation algorithms," *Appl. Acoust.*, vol. 188, Jan. 2022, Art. no. 108539.
[110] S. Basir, M. N. Hossain, M. S. Hosen, M. A. Rahman, M. Aktaruzzaman, Ahsan-Ul-Ambia, and M. S. Islam. *Supervised Single Channel Source Separation Using U-Net*. Accessed: Dec. 20, 2023. [Online]. Available: https://ssrn.com/abstract=4423737
[111] C. Pang, J. Fan, Q. Shen, Y. Xie, C. Huang, and B. W. Schuller, "Multichannel speech enhancement based on neural beamforming and a context-focused post-filtering network," *IEEE Trans. Cogn. Develop. Syst.*, early access, Sep. 2023.
[112] R. Gao and K. Grauman, "VisualVoice: Audio-visual speech separation with cross-modal consistency," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 15490–15500.
[113] Y. Wu, C. Li, and Y. Qian, "Light-weight VisualVoice: Neural network quantization on audio visual speech separation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. Workshops (ICASSPW)*, Jun. 2023, pp. 1–5.
[114] M. Yoshida, R. Togo, T. Ogawa, and M. Haseyama, "Off-screen sound separation based on audio-visual pre-training using binaural audio," *Sensors*, vol. 23, no. 9, p. 4540, May 2023.
[115] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 7132–7141.
[116] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," 2015, *arXiv:1503.03832v3*.
[117] R. M. Bittner, J. Salamon, M. Tierney, M. Mauch, C. Cannam, and J. P. Bello, "MedleyDB: A multitrack dataset for annotationintensive MIR research," in *Proc. 15th Int. Soc. Music Inf. Retr. Conf. (ISMIR)*, Taipei, Taiwan, Oct. 2014, pp. 155–160.
[118] T.-S. Chan, T.-C. Yeh, Z.-C. Fan, H.-W. Chen, L. Su, Y.-H. Yang, and R. Jang, "Vocal activity informed singing voice separation with the iKala dataset," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2015, pp. 718–722.
[119] *Demixing Secret Data*. Accessed: Dec. 20, 2023. [Online]. Available: https://sigsep.github.io/datasets/dsd100.html
[120] A. Moryossef, Y. Elazar, and Y. Goldberg, "At your fingertips: Automatic piano fingering detection," in *Proc. ICLR Conf.*, Sep. 2019, pp. 1–11.
[121] R. Gao, R. Feris, and K. Grauman, "Learning to separate object sounds by watching unlabeled video," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 35–53.
[122] Y. Tian, J. Shi, B. Li, Z. Duan, and C. Xu, "Audio-visual event localization in unconstrained videos," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 247–263.
[123] X. Hu, Z. Chen, and A. Owens, "Mix and localize: Localizing sound sources in mixtures," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 10473–10482.
[124] H. Chen, W. Xie, A. Vedaldi, and A. Zisserman, "A large-scale audio-visual dataset," in *IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, pp. 721–725, 2020.
[125] J. Carreira and A. Zisserman, "Quo vadis, action recognition? A new model and the kinetics dataset," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6299–6308.
[126] M. Fukushima, A. Doyle, M. Mallarkey, M. Mishkin, and B. Averbeck, "Data from: Distributed acoustic cues for caller identity in macaque vocalization," in *Roy. Soc. open Sci.*, vol. 2, no. 12, 2015, Art. no. 150432.
[127] L. S. Sayigh, H. C. Esch, R. S. Wells, and V. M. Janik, "Facts about signature whistles of bottlenose dolphins, tursiops truncatus," *Animal Behaviour*, vol. 74, no. 6, pp. 1631–1642, Dec. 2007.
[128] Y. Prat, M. Taub, E. Pratt, and Y. Yovel, "An annotated dataset of Egyptian fruit bat vocalizations across varying contexts and during vocal ontogeny," *Sci. Data*, vol. 4, no. 1, Oct. 2017, Art. no. 170143.
[129] S. R. Ness. (Oct. 2021). *Orchive*. [Online]. Available: http://orchive.cs.uvic.ca/
[130] W. Cukierski, A. Karpištšenko, and E. Spaulding. (2013). *The Marinexplore and Cornell University Whale Detection Challenge*. Accessed: Dec. 20, 2023. [Online]. Available: https://kaggle.com/competitions/whale-detection-challenge
[131] *TIMIT Dataset*. Accessed: Dec. 20, 2023. [Online]. Available: https://catalog.ldc.upenn.edu/LDC93s1
[132] *McGill Dataset*. Accessed: Dec. 20, 2023. [Online]. Available: https://www.mmsp.ece.mcgill.ca/Documents/Data/
[133] C. Richey, M. A. Barrios, Z. Armstrong, C. Bartels, H. Franco, M. Graciarena, A. Lawson, M. K. Nandwana, A. Stauffer, J. van Hout, P. Gamble, J. Hetherly, C. Stephenson, and K. Ni, "Voices obscured in complex environmental settings (voices) corpus," 2018, *arXiv:1804.05053*.
[134] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third 'CHiME' speech separation and recognition challenge: Dataset, task and baselines," in *Proc. IEEE Workshop Autom. Speech Recognit. Understand. (ASRU)*, Dec. 2015, pp. 504–511.
[135] R. M. Corey, N. Tsuda, and A. C. Singer, "Acoustic impulse responses for wearable audio devices," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 216–220.
[136] J. S. Chung, A. Nagrani, and A. Zisserman, "VoxCeleb2: Deep speaker recognition," Jun. 2018, *arXiv:1806.05622*.
[137] J.-C. Hou, S.-S. Wang, Y.-H. Lai, Y. Tsao, H.-W. Chang, and H.-M. 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.
[138] N. Harte and E. Gillen, "TCD-TIMIT: An audio-visual corpus of continuous speech," *IEEE Trans. Multimedia*, vol. 17, no. 5, pp. 603–615, May 2015.
[139] E. K. Patterson et al., "CUAVE: A new audio-visual database for multimodal human–computer interface research," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, May 2002, Paper II-2017.
[140] T. Afouras, J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Deep audio-visual speech recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 12, pp. 8717–8727, Dec. 2022.
[141] A. Nagrani, J. S. Chung, and A. Zisserman, "VoxCeleb: A large-scale speaker identification dataset," Jun. 2017, *arXiv:1706.08612*.

[142] R. Gao and K. Grauman, "2.5D visual sound," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 324–333.

[143] *BBC Dataset*. Accessed: Dec. 20, 2023. [Online]. Available: https://www.robots.ox.ac.uk/vgg/data/lip_reading/lrs2.html

[144] C. Veaux, J. Yamagishi, and S. King, "The voice bank corpus: Design, collection and data analysis of a large regional accent speech database," in *Proc. Int. Conf. Oriental COCOSDA Conf. Asian Spoken Lang. Res. Eval. (O-COCOSDA/CASLRE)*, Nov. 2013, pp. 1–4.

[145] J. Thiemann, N. Ito, and E. Vincent, "The diverse environments multi-channel acoustic noise database: A database of multichannel environmental noise recordings," *J. Acoust. Soc. Amer.*, vol. 133, no. 5, p. 3591, May 2013.

[146] D. Yin, C. Luo, Z. Xiong, and W. Zeng, "Phasen: A phase-and-harmonics-aware speech enhancement network," in *Proc. AAAI Conf. Artif. Intell.*, Apr. 2020, vol. 34, no. 5, pp. 9458–9465.

[147] A. Marafioti, N. Perraudin, N. Holighaus, and P. Majdak, "A context encoder for audio inpainting," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 12, pp. 2362–2372, Dec. 2019.

[148] P. J. Wolfe and S. J. Godsill, "Interpolation of missing data values for audio signal restoration using a Gabor regression model," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, vol. 5, Apr. 2005, pp. 517–520.

[149] L. Oudre, "Interpolation of missing samples in sound signals based on autoregressive modeling," *Image Process. Line*, vol. 8, pp. 329–344, Oct. 2018.

[150] M. Lagrange, S. Marchand, and J.-B. Rault, "Long interpolation of audio signals using linear prediction in sinusoidal modeling," *J. Audio Eng. Soc.*, vol. 53, no. 10, pp. 891–905, 2005.

[151] M. Kegler, P. Beckmann, and M. Cernak, "Deep speech inpainting of time-frequency masks," in *Proc. Interspeech*, Oct. 2020.

[152] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 770–778.

[153] C. Aironi, S. Cornell, L. Serafini, and S. Squartini, "A time-frequency generative adversarial based method for audio packet loss concealment," in *Proc. 31st Eur. Signal Process. Conf. (EUSIPCO)*, Sep. 2023, pp. 121–125.

[154] *Pix2Pix Code*. Accessed: Dec. 20, 2023. [Online]. Available: https://www.tensorflow.org/tutorials/generative/pix2pix

[155] D. Griffin and J. Lim, "Signal estimation from modified short-time Fourier transform," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-32, no. 2, pp. 236–243, Apr. 1984.

[156] H. Zhao, "A GAN speech inpainting model for audio editing software," in *Proc. Interspeech*, Aug. 2023, pp. 5127–5131.

[157] Z. Pruša, P. Balazs, and P. L. Søndergaard, "A noniterative method for reconstruction of phase from STFT magnitude," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 5, pp. 1154–1164, May 2017.

[158] N. Perraudin, N. Holighaus, P. Majdak, and P. Balazs, "Inpainting of long audio segments with similarity graphs," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 6, pp. 1083–1094, Jun. 2018.

[159] Y. Li, B. Gfeller, M. Tagliasacchi, and D. Roblek, "Learning to denoise historical music," 2020, arXiv:2008.02027.

[160] 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)*, Singapore, May 2022, pp. 841–845.

[161] I. Irigaray et al., "Noise reduction in analog tape audio recordings with deep learning models," in *Proc. Int. Conf. Audio Archiving, Preservation Restoration (AES)*, Virginia, USA, Jun. 2023, pp. 1–6.

[162] E. Moliner, J. Lehtinen, and V. Välimäki, "Solving audio inverse problems with a diffusion model," in *Proc. ICASSP - IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2023, pp. 1–5.

[163] E. Moliner and V. Välimäki, "Diffusion-based audio inpainting," 2023, *arXiv:2305.15266*.

[164] P. Warden, "Speech commands: A dataset for limited-vocabulary speech recognition," 2018, *arXiv:1804.03209*.

[165] K. J. Piczak, "ESC: Dataset for environmental sound classification," in *Proc. 23rd ACM Int. Conf. Multimedia*, Oct. 2015, pp. 1015–1018.

[166] J. Yamagishi, C. Veaux, and K. MacDonald, "CSTR VCTK Corpus: English multi-speaker corpus for CSTR voice cloning toolkit (version 0.92)," Centre Speech Technol. Res. (CSTR), Univ. Edinburgh, Edinburgh, U.K., 2019.

[167] C. Valentini-Botinhao, X. Wang, S. Takaki, and J. Yamagishi, "Investigating RNN-based speech enhancement methods for noise-robust text-to-speech," in *Proc. 9th ISCA Workshop Speech Synth. Workshop (SSW)*, Sep. 2016, pp. 146–152.

[168] *Public Domain Project Dataset*. Accessed: Dec. 20, 2023. [Online]. Available: http://pool.publicdomainproject.org

[169] *The Great 78 Project Dataset*. Accessed: Dec. 20, 2023. [Online]. Available: https://great78.archive.org

[170] J. Thickstun, Z. Harchaoui, and S. M. Kakade, "Learning features of music from scratch," in *Proc. ICLR*, Toulon, France, Apr. 2017. Accessed: Dec. 20, 2023. [Online]. Available: https://www.kaggle.com/datasets/imsparsh/musicnet-dataset

[171] The Internet Archive. *Orchestral and Opera Recordings*. Accessed: Dec. 20, 2023. [Online]. Available: https://archive.org

[172] C. Hawthorne, A. Stasyuk, A. Roberts, I. Simon, C.-Z. A. Huang, S. Dieleman, E. Elsen, J. Engel, and D. Eck, "Enabling factorized piano music modeling and generation with the MAESTRO dataset," in *Proc. ICLR*, May 2019.

[173] M. I. Mandel, R. J. Weiss, and D. P. W. Ellis, "Model-based expectation-maximization source separation and localization," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 2, pp. 382–394, Feb. 2010.

[174] A. E. Bulut and K. Koishida, "Low-latency single channel speech enhancement using U-Net convolutional neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Barcelona, Spain, May 2020, pp. 6214–6218.

[175] G. W. Lee, K. M. Jeon, and H. K. Kim, "U-Net-based single-channel wind noise reduction in outdoor environments," in *Proc. IEEE Int. Conf. Consum. Electron. (ICCE)*, Jan. 2020, pp. 1–2.

[176] Y. J. Lee, Y. H. Moon, J. Y. Park, and O. G. Min, "Recent R&D trends for lightweight deep learning," *Electrics Telecommun. Trends*, vol. 34, no. 2, pp. 40–50, 2019.

[177] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, Inception-ResNet and the impact of residual connections on learning," in *Proc. 31st AAAI Conf. Artif. Intell.*, vol. 31, no. 1, CA, USA, Feb. 2017, pp. 4278–4284.

[178] K. Akter, N. Mamun, and Md. A. Hossain, "A T-F masking based monaural speech enhancement using U-Net architecture," in *Proc. Int. Conf. Electr., Comput. Commun. Eng. (ECCE)*, Chittagong, Bangladesh, Feb. 2023, pp. 1–5.

[179] H. Chung, V. S. Tomar, and B. Champagne, "Deep convolutional neural network-based inverse filtering approach for speech de-reverberation," 2020, *arXiv:2010.07895v1*.

[180] S. Gul, M. S. Khan, and S. W. Shah, "Preserving the beamforming effect for spatial cue-based pseudo-binaural dereverberation of a single source," *Comput. Speech Lang.*, vol. 77, Jan. 2023, Art. no. 101445.

[181] I. Tashev and H. Malvar, "System and method for beamforming using a microphone array," U.S. Patent U.S. 7 415 117 B2, Aug. 2008.

[182] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *JASA*, vol. 66, no. 4, pp. 943–950, Apr. 1979.

[183] T. Yoshioka and T. Nakatani, "Generalization of multi-channel linear prediction methods for blind MIMO impulse response shortening," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 10, pp. 2707–2720, Dec. 2012.

[184] Z.-Q. Wang and D. Wang, "Deep learning based target cancellation for speech dereverberation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 941–950, 2020.

[185] E. J. Nustede and J. Anemüller, "Single-channel speech enhancement with deep complex U-networks and probabilistic latent space models," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Rhodes Island, Greece, Jun. 2023, pp. 1–5.

[186] K. M. Jeon, G. W. Lee, N. K. Kim, and H. K. Kim, "TAU-Net: Temporal activation U-Net shared with nonnegative matrix factorization for speech enhancement in unseen noise environments," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 3400–3414, 2021.

[187] P. Dhariwal and A. Nichol, "Diffusion models beat GANs on image synthesis," in *Proc. Adv. Neural Inf. Proc. Syst. (NeurIPS)*, vol. 34, Dec. 2021, pp. 8780–8794.

[188] T. B. Brown et al., "Language models are few-shot learners," 2020, *arXiv:2005.14165*.

[189] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "WaveNet: A generative model for raw audio," 2016, *arXiv:1609.03499*.

[190] P. Dhariwal, H. Jun, C. Payne, J. W. Kim, A. Radford, and I. Sutskever, "Jukebox: A generative model for music," 2020, *arXiv:2005.00341*.
[191] S. Liu, Y. Cao, D. Su, and H. Meng, "DiffSVC: A diffusion probabilistic model for singing voice conversion," in *Proc. IEEE Autom. Speech Recognit. Understand. Workshop (ASRU)*, Dec. 2021, pp. 741–748.
[192] H. Yen, F. G. Germain, G. Wichern, and J. L. Roux, "Cold diffusion for speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Rhodes Island, Greece, Jun. 2023, pp. 1–5.
[193] Y.-J. Lu, Y. Tsao, and S. Watanabe, "A study on speech enhancement based on diffusion probabilistic model," in *Proc. Asia–Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA ASC)*, Dec. 2021, pp. 659–666.
[194] Y.-J. Lu, Z.-Q. Wang, S. Watanabe, A. Richard, C. Yu, and Y. Tsao, "Conditional diffusion probabilistic model for speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Singapore, May 2022, pp. 7402–7406.
[195] S. Welker, J. Richter, and T. Gerkmann, "Speech enhancement with score-based generative models in the complex STFT domain," 2022, *arXiv:2203.17004*.
[196] J. Richter, S. Welker, J.-M. Lemercier, B. Lay, and T. Gerkmann, "Speech enhancement and dereverberation with diffusion-based generative models," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 31, pp. 2351–2364, 2023.
[197] Y. Hu, Y. Liu, S. Lv, M. Xing, S. Zhang, Y. Fu, J. Wu, B. Zhang, and L. Xie, "DCCRN: Deep complex convolution recurrent network for phase-aware speech enhancement," 2020, *arXiv:2008.00264*.
[198] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2021.
[199] S. Y. Lee, J. Chang, and S. Lee, "Deep learning-based method for multiple sound source localization with high resolution and accuracy," *Mech. Syst. Signal Process.*, vol. 161, Dec. 2021, Art. no. 107959.
[200] T. S. Sharan, R. Bhattacharjee, S. Sharma, and N. Sharma, "Evaluation of deep learning methods (DnCNN and U-Net) for denoising of heart auscultation signals," in *Proc. 3rd Int. Conf. Commun. Syst., Comput. IT Appl. (CSCITA)*, Mumbai, India, Apr. 2020, pp. 151–155.
[201] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, "Beyond a Gaussian denoiser: Residual learning of deep CNN for image denoising," *IEEE Trans. Image Process.*, vol. 26, no. 7, pp. 3142–3155, Jul. 2017.
[202] A. Mukherjee, R. Banerjee, and A. Ghose, "A novel U-Net architecture for denoising of real-world noise corrupted phonocardiogram signal," 2023, *arXiv:2310.00216*.
[203] C. González–rodríguez, M. A. Alonso–arévalo, and E. García–canseco, "Robust denoising of phonocardiogram signals using time-frequency analysis and U-Nets," *IEEE Access*, vol. 11, pp. 52466–52479, 2023.
[204] S. Liu, A. Mallol-Ragolta, and B. W. Schuller, "COVID-19 detection from speech in noisy conditions," in *Proc. ICASSP - IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Rhodes Island, Greece, Jun. 2023, pp. 1–5.
[205] T. S. Jahren, A. R. Sørnes, B. Dénarié, E. Steen, T. Bjåstad, and A. H. S. Solberg, "Reverberation suppression in echocardiography using a causal convolutional neural network," *IEEE Access*, vol. 11, pp. 67922–67937, 2023.
[206] J. Yamashita et al., "Anomaly detection using autoencoder, IDNN and U-Net using ensemble," in *Proc. Challenge Detection Classification Acoustic Scenes Events (DCASE)*, 2022, pp. 1–5.
[207] *DCASE 2018 Task 1 and 2 Dataset*. Accessed: Dec. 18, 2023. [Online]. Available: https://dcase.community/challenge2022/index
[208] P. Daniluk et al., "Ensemble of auto-encoder based and WaveNet like systems for unsupervised anomaly detection," in *Proc. Challenge Detection Classification Acoustic Scenes Events (DCASE)*, 2020, pp. 1–5.
[209] Y. Shin, Y. G. Kim, C.-H. Choi, D.-J. Kim, and C. Chun, "SELD U-Net: Joint optimization of sound event localization and detection with noise reduction," *IEEE Access*, vol. 11, pp. 105379–105393, Sep. 2023.
[210] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: A nested U-Net architecture for medical image segmentation," in *Proc. Deep Learn. Med. Image Anal. Multimodal Learn. Clin. Decis. Support*, 2018, pp. 3–11.
[211] P. Zhang and C. Chen, "A two-stage framework for time-frequency analysis and fault diagnosis of planetary gearboxes," *Appl. Sci.*, vol. 13, no. 8, p. 5202, Apr. 2023.
[212] P. Zhang and C. Chen, "Time–frequency analysis for planetary gearbox fault diagnosis based on improved U-Net++," *J. Failure Anal. Prevention*, vol. 23, no. 3, pp. 1068–1080, Jun. 2023.
[213] S. S. Salehi, D. Erdogmus, and A. Gholipour, "Tversky loss function for image segmentation using 3D fully convolutional deep networks," in *Proc. Int. Workshop Mach. Learn. Med. Imag.* Cham, Switzerland: Springer, Sep. 2017, pp. 379–387.
[214] Y. Sudo, K. Itoyama, K. Nishida, and K. Nakadai, "Environmental sound segmentation utilizing mask U-Net," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Nov. 2019, pp. 5340–5345.
[215] A. Mesaros, T. Heittola, and T. Virtanen, "A multi-device dataset for urban acoustic scene classification," in *Proc. Detection Classification Acoustic Scenes Events Workshop (DCASE)*, Surrey, U.K., Nov. 2018, pp. 9–13.
[216] E. Fonseca, M. Plakal, F. Font, D. P. W. Ellis, X. Favory, J. Pons, and X. Serra, "General-purpose tagging of freesound audio with audioset labels: Task description, dataset, and baseline," in *Proc. Detection Classification Acoustic Scenes Events Workshop (DCASE)*, Surrey, U.K., Nov. 2018, pp. 69–73.
[217] B.-J. Lee, M.-S. Lee, and W.-S. Jung, "Acoustic based fire event detection system in underground utility tunnels," *Fire*, vol. 6, no. 5, p. 211, May 2023.
[218] A. Larcher, K. A. Lee, B. Ma, and H. Li, "Text-dependent speaker verification: Classifiers, databases and RSR2015," *Speech Commun.*, vol. 60, pp. 56–77, May 2014.
[219] Audio Labs. *RIR Generator*. Accessed: Dec. 20, 2023. [Online]. Available: https://www.audiolabs-erlan-en.de/fau/professor/habets/software/rir-generator/
[220] E. A. P. Habets, "Room impulse response generator," Technische Universiteit Eindhoven, Eindhoven, The Netherlands, Tech. Rep. 2(2.4), 2006, vol. 1.
[221] R. Stewart and M. Sandler, "Database of omnidirectional and B-format room impulse responses," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Mar. 2010, pp. 165–168.
[222] C. K. A. Reddy, E. Beyrami, J. Pool, R. Cutler, S. Srinivasan, and J. Gehrke, "A scalable noisy speech dataset and online subjective test framework," in *Proc. Interspeech*, Sep. 2019, pp. 1816–1820.
[223] K. Kinoshita, M. Delcroix, S. Gannot, E. A. P. Habets, R. Haeb-Umbach, W. Kellermann, V. Leutnant, R. Maas, T. Nakatani, B. Raj, A. Sehr, and T. Yoshioka, "A summary of the REVERB challenge: State-of-the-art and remaining challenges in reverberant speech processing research," *EURASIP J. Adv. Signal Process.*, vol. 2016, no. 1, pp. 1–19, Dec. 2016.
[224] *Surrey RiRs*. Accessed: Dec. 20, 2023. [Online]. Available: https://openresearch.surrey.ac.uk/esploro/outputs/dataset/S3A-Room-Impulse-Responses/99513590402346
[225] T. Grzywalski and S. Drgas, "Speech enhancement using U-Nets with wide-context units," *Multimedia Tools Appl.*, vol. 81, no. 13, pp. 18617–18639, May 2022.
[226] J. S. Garofolo, D. Graff, D. Paul, and D. Pallett, 2007, "CSR-I (WSJ0) complete," *Linguistic Data Consortium*, doi: 10.35111/ewkm-cg47.
[227] *Freesound Dataset*. Accessed: Dec. 20, 2023. [Online]. Available: https://freesound.org/
[228] A. Varga and H. J. M. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Commun.*, vol. 12, no. 3, pp. 247–251, Jul. 1993.
[229] C. Liu et al., "An open access database for the evaluation of heart sound algorithms," *Physiol. Meas.*, vol. 37, no. 12, pp. 2181–2213, Dec. 2016.
[230] P. Bentley et al. *The PASCAL Classifying Heart Sounds Challenge 2011 (CHSC2011) Results*. Accessed: Dec. 20, 2023. [Online]. Available: http://www.peterjbentley.com/heartchallenge/index.html
[231] Yaseen, G.-Y. Son, and S. Kwon, "Classification of heart sound signal using multiple features," *Appl. Sci.*, vol. 8, no. 12, p. 2344, Nov. 2018. [Online]. Available: https://www.mdpi.com/2076-3417/8/12/2344
[232] N. K. Sharma, S. R. Chetupalli, D. Bhattacharya, D. Dutta, P. Mote, and S. Ganapathy, "The second dicova challenge: Dataset and performance analysis for diagnosis of covid-19 using acoustics," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2022, pp. 556–560.
[233] *DCASE 2022 Task 2 Dataset*. Accessed: Dec. 18, 2023. [Online]. Available: https://dcase.community/challenge2020/index
[234] *ATRECSS Dataset*. Accessed: Dec. 20, 2023. [Online]. Available: http://www.atr.jp

[235] *RWCP Sound Scene Database*. Accessed: Dec. 20, 2023. [Online]. Available: https://www.openslr.org/13/
[236] *Freesound General-Purpose Audio Tagging Challenge Dataset*. Accessed: Dec. 20, 2023. [Online]. Available: https://www.kaggle.com/c/freesound-audio-tagging
[237] *DCASE 2016 Dataset*. Accessed: Dec. 20, 2023. [Online]. Available: https://dcase.community/challenge2016/task-sound-event-detection-in-synthetic-audio-results
[238] *RWC-Music Database*. Accessed: Dec. 20, 2023. [Online]. Available: https://staff.aist.go.jp/m.goto/RWC-MDB/
[239] *Xeno-Canto Dataset*. Accessed: Dec. 20, 2023. [Online]. Available: https://xeno-canto.org/explore
[240] S. B. Shuvo, S. N. Ali, S. I. Swapnil, T. Hasan, and M. I. H. Bhuiyan, "A lightweight CNN model for detecting respiratory diseases from lung auscultation sounds using EMD-CWT-based hybrid scalogram," *IEEE J. Biomed. Health Informat.*, vol. 25, no. 7, pp. 2595–2603, Jul. 2021.
[241] M. Lech, M. Stolar, R. Bolia, and M. Skinner, "Amplitude-frequency analysis of emotional speech using transfer learning and classification of spectrogram images," *Adv. Sci., Technol. Eng. Syst. J.*, vol. 3, no. 4, pp. 363–371, Aug. 2018.
[242] C. Lin, "Face detection in complicated backgrounds and different illumination conditions by using YCbCr color space and neural network," *Pattern Recognit. Lett.*, vol. 28, no. 16, pp. 2190–2200, Dec. 2007.
[243] B. Sreedha, P. R. Nair, and R. Maity, "Non-invasive early diagnosis of jaundice with computer vision," *Proc. Comput. Sci.*, vol. 218, pp. 1321–1334, Jan. 2023.
[244] H. Kim, H. Lee, S. Ahn, W.-K. Jung, and S.-H. Ahn, "Broken stitch detection system for industrial sewing machines using HSV color space and image processing techniques," *J. Comput. Design Eng.*, vol. 10, no. 4, pp. 1602–1614, Jul. 2023.
[245] H. Zhao, J. Jin, Y. Liu, Y. Guo, and Y. Shen, "FSDF: A high-performance fire detection framework," *Expert Syst. Appl.*, vol. 238, Mar. 2024, Art. no. 121665.
[246] T. G. Devi, N. Patil, S. Rai, and C. S. Philipose, "Gaussian blurring technique for detecting and classifying acute lymphoblastic leukemia cancer cells from microscopic biopsy images," *Life*, vol. 13, no. 2, p. 348, Jan. 2023.
[247] P. Appiahene, E. J. Arthur, S. Korankye, S. Afrifa, J. W. Asare, and E. T. Donkoh, "Detection of anemia using conjunctiva images: A smartphone application approach," *Med. Novel Technol. Devices*, vol. 18, Jun. 2023, Art. no. 100237.
[248] S. Gul, M. S. Khan, and M. Fazeel, "Single channel speech enhancement by colored spectrograms," 2023, *arXiv:2310.17142*.
[249] Y. Deng, Y. Hou, J. Yan, and D. Zeng, "ELU-Net: An efficient and lightweight U-Net for medical image segmentation," *IEEE Access*, vol. 10, pp. 35932–35941, 2022.
[250] J. Zhang, H. Zhu, P. Wang, and X. Ling, "ATT squeeze U-Net: A lightweight network for forest fire detection and recognition," *IEEE Access*, vol. 9, pp. 10858–10870, 2021.
[251] Y. Li, E. Chouzenoux, B. Charmettant, B. Benatsou, J.-P. Lamarque, and N. Lassau, "Lightweight U-Net for lesion segmentation in ultrasound images," in *Proc. IEEE 18th Int. Symp. Biomed. Imag. (ISBI)*, Nice, France, Apr. 2021, pp. 611–615.
[252] J. Duan, X. Liu, X. Wu, and C. Mao, "Detection and segmentation of iron ore green pellets in images using lightweight U-Net deep learning network," *Neural Comput. Appl.*, vol. 32, no. 10, pp. 5775–5790, May 2020.
[253] *MATLAB Spectrogram*. Accessed: Dec. 20, 2023. [Online]. Available: https://www.mathworks.com/help/signal/ref/spectrogram.html
[254] *MATLAB Mel-Spectrogram*. Accessed: Dec. 20, 2023. [Online]. Available: https://www.mathworks.com/help/audio/ref/melspectrogram.html
[255] *Time-Frequency Analysis and Continuous Wavelet Transform*. Accessed: Dec. 20, 2023. [Online]. Available: https://ww2.mathworks.cn/help/wavelet/ug/time-frequency-analysis-and-continuous-wavelet-transform.html

SANIA GUL received the B.Sc., M.Sc., and Ph.D. degrees in electrical engineering from the University of Engineering and Technology, Peshawar, Peshawar, Pakistan, in 2001, 2006, and 2021, respectively. Her research interests include audio processing, 6G networks, machine learning, and deep learning.

MUHAMMAD SALMAN KHAN received the M.S. degree in electrical engineering from The George Washington University, Washington, DC, USA, in 2010, and the Ph.D. degree in electrical engineering from Loughborough University, U.K., in 2013. He was a Postdoctoral Fellow with the Department of Electrical Engineering, Universidad de Chile, Santiago, Chile, from 2013 to 2015. He was the Principal Investigator of the Higher Education Commission Pakistan funded Intelligent Information Processing Laboratory, National Center of Artificial Intelligence, from 2017 to 2021. In 2019, he was included as a member of the World Health Organization (WHO) Roster of Experts on Digital Health. He is currently an Assistant Professor with the Department of Electrical Engineering, College of Engineering, Qatar University, Doha, Qatar. His research interests include signal processing, pattern recognition, machine learning, and artificial intelligence. He is a member of the IEEE Signal Processing Society.