IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 14 July 2024, accepted 1 August 2024, date of publication 5 August 2024, date of current version 16 August 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3438947

RESEARCH ARTICLE

Feature Aggregation in Joint Sound Classification and Localization Neural Networks

BRENDAN HEALY, PATRICK MCNAMEE, AND ZAHRA NILI AHMADABADI, (Member, IEEE)
Department of Mechanical Engineering, San Diego State University, San Diego, CA 92182, USA
Corresponding author: Zahra Nili Ahmadabadi (zniliahmadabadi@sdsu.edu)

This work was supported by the San Diego State University Seed Grant Program.

ABSTRACT Current state-of-the-art sound source localization (SSL) deep learning networks lack feature aggregation within their architecture. Feature aggregation within neural network architectures enhances model performance by enabling the consolidation of information from different feature scales, thereby improving feature robustness and invariance. We adapt feature aggregation sub-architectures from computer vision neural networks onto a baseline neural network architecture for SSL, the Sound Event Localization and Detection network (SELDnet). The incorporated sub-architecture are: Path Aggregation Network (PANet); Weighted Bi-directional Feature Pyramid Network (BiFPN); and a novel Scale Encoding Network (SEN). These sub-architectures were evaluated using two metrics for signal classification and two metrics for direction-of-arrival regression. The results show that models incorporating feature aggregations outperformed the baseline SELDnet, in both sound signal classification and localization. Among the feature aggregators, PANet exhibited superior performance compared to other methods, which were otherwise comparable. The results provide evidence that feature aggregation sub-architectures enhance the performance of sound detection neural networks, particularly in direction-of-arrival regression.

INDEX TERMS Joint sound signal classification and localization, multi-task deep learning, feature aggregation.

I. INTRODUCTION

Sound source localization (SSL) represents an imperative domain within the broader field of audio signal processing, holding significant implications for topics such as robotics, hearing aids, and speech recognition systems [1]. SSL techniques aim to ascertain the location and/or direction-of-arrival (DOA) of a sound source, which provides critical data for sound source separation [2], speech augmentation [3], robot-human interaction [4], noise control [5], and auditory scene analysis [6]. A key gap within existing SSL neural networks is the lack of feature aggregation within their architecture. Feature aggregation can boost a model performance by consolidating information from various scales and contexts, thereby enhancing feature robustness and scale invariance. It is particularly vital for SSL networks, which must distinguish between direct signals and reflections [7]. We adapt feature aggregation techniques from computer vision neural networks and applying them to signal detection neural networks. Additionally, we propose a novel architecture, the Scale Encoding Network (SEN), which serves as a compact feature aggregator in the context of SSL.

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

A. RELATED WORK

Early endeavors in machine learning for SSL were focused on conventional machine learning models, namely the Multilayer Perceptron (MLP) and Support Vector Machines (SVM) [8], [9]. The aforementioned models encountered difficulties, particularly in effectively managing large datasets and addressing the complexities associated with temporal relationships in the input features. In light of these difficulties, there has been a notable shift towards Convolutional Neural Networks (CNNs), which have demonstrated the ability to capture spatial features in data [6], [10]. As deep learning techniques have advanced, the development of Recurrent Neural Networks (RNNs) gave rise to Convolutional Recurrent Neural Networks (CRNNs). This combination

successfully utilized both spatial and temporal dimensions of the data, leading to improved DOA estimation [6], [11], [12].

Residual CNNs (Res-CNNs) soon emerged, incorporating shortcut connections that provide a link between the input and output layers. Res-CNNs proved superior in performance compared to both conventional CNNs and CRNNs [13], [14]. The introduction of novel architectures such as Res-CRNNs and deep generative models enhanced the DOA estimation [15], [16], [17]. Additionally, attention mechanisms have continued to enhance the capabilities of neural networks by allowing them to selectively concentrate on pertinent features, improving the accuracy of the estimation process [18], [19], [20], [21], [22].

In addition to single-task SSL networks, neural networks have been applied to the multi-task problem of Sound Event Localization and Detection (SELD), where there are multiple sound classes to be detected and localized at once. Hirvonen [23] was the first to solve the SELD problem by treating it as a multi-class classification task using a CNN. In 2018, Adavanne et al. [7] introduced the first CRNN-based SELD method (SELDnet) which was effective in scenes with more than two overlapping sound events and was able to localize sources at any azimuth and elevation angles. SELDnet used a single network but with two branches dedicated to solve the Sound Event Detection (SED) problem and the SSL problem. The input features to SELDnet included magnitude and phase of spectrograms. After this study, SELD gained overwhelming interest from the community, leading to numerous advancements aimed at improving its performance. Researchers proposed new deep learning-based networks [24], [25], [26], [27], [28], [29], [30], [31], [32], [33], [34], [35], [36], [37], [38], explored novel input features [39], [40], and developed innovative data augmentation techniques [26]. Some of the most significant studies are as follows. Cao et al. [24] proposed using two networks, instead of a single network as in SELDnet, to better balance the SED and SSL objectives during optimization. However, this representation increased system complexity and network size. To address, Shimada et al. [25] proposed an activity-coupled Cartesian DOA (ACCDOA) representation, which avoids increasing model size. Additionally, [41] presented an Event-Independent Network that enabled detection of different sound events of the same type but with different DOAs. Wang et al. [26] introduced a four-stage spatial data augmentation approach to increase the diversity of DOA representations in limited training datasets while effectively dealing with overlapping sources. They also formed a ResNet-Conformer architecture to capture both global and local context dependencies in an audio sequence. While conventional SELD methods only consider static microphones, a recent study by [42] formulated a multi-modal SELD that utilizes both the audio and motion tracking sensor signals.

The above-discussed deep learning-based SSL and SELD have used datasets collected with different configurations of microphone arrays [43]. The number of microphones in the array is determined based on various factors [44], including human imitation (e.g., using 2-microphone array systems), the aim to reduce the dataset size and computational time, and the argument that SSL performance is directly correlated to the number of microphones. The existing studies have used arrays containing varying numbers of microphones, including 2 (also termed binaural arrays) [45], [46], [47], [48], [49], [50], [51], [52], [53], [54], [55], [56], 4 [57], [58], 8 [59], or more [7], [60]. The features used as inputs to the network also vary significantly [6], involving different levels of pre-processing of the datasets. This ranges from no pre-processing, when using raw multichannel waveforms, to higher levels of pre-processing for networks using interchannel features [61], [62], cross correlation-based features [63], [64], or intensity-based features [41], [65]. A commonly used type of input data across studies involves the extraction of spectrogram-based features, which require minimal pre-processing, namely computing the Short-Time Fourier Transform (STFT) of multichannel information. Various combinations of these spectrogram-based features have been employed. Some studies have omitted either phase [66], [67] or magnitude information [68], while others have chosen to retain both [69], [70], arguing that keeping both magnitude and phase (or alternatively real and imaginary parts) can improve localization performance.

Lastly, feature aggregation has been previously used in neural networks for SELD and sound classification. However, these efforts have mainly focused on aggregations of features into the inputs of the neural network, rather than a change in the feature architecture. Some of these features are statistical features of the audio [71] while others are learned through dictionary methods via clustering [72]. The closest previous work to achieving sound classification with feature aggregation as part of the network architecture is [73] which involved audio tagging for music genre classification. The network in [73] directly feed in convolutional neural network layers into a fully connected network for tag prediction. This architecture therefore has a backbone and head architecture sections but fails to incorporate the neck section prevalent in the PANet or BiFPN. These approaches may not fully leverage the hierarchical nature of the neural networks and tend to be less adaptive in processing features. The present study formulates methods to incorporate feature aggregation in the neck of the SELD network to address these shortcomings and enable the network to combine features from the different scales of the backbone.

B. CONTRIBUTIONS

This paper makes the following contributions:

1) We introduce Feature Aggregation techniques from image based Object Detection Neural Networks into SSL and Sound Detection Res-CRNN networks.
2) We propose and test a new Feature Aggregator method, the SEN.

3) We provide a publicly available Feature Aggregator Library for TensorFlow's Functional API. This library contains pre-made aggregators and allows for the efficient creation of new aggregators.

C. OVERVIEW

The remainder of this paper is organized as follows. Section II explains challenges with feature scaling in SSL networks, the role of feature aggregation in resolving them, and its practical application in real-world models. Section III elaborates on the training and testing procedures used to gather data for evaluating the merit of our theory. Section IV describes the dataset, preprocessing, evaluation metrics, and baseline methods used to gauge our results. Section V investigates the testing outcomes and their contextual meaning. Section VI provides a summary of our findings, their implications, and directions for future research.

II. THEORY

This section is divided into four parts. Section II-A discusses the importance of scale invariance within neural networks. Section II-B explains how feature aggregation addresses scale invariance, how feature aggregation is performed, and various feature aggregation designs. Section II-C elaborates on our custom aggregation approach. Section II-D gives an overview of standard object detection designs and how feature aggregation is employed within a full architecture.

A. SCALING

Feature scaling is a powerful tool in both object detection and SSL neural networks. The ability to learn notable patterns in data and then identify these patterns at different scales reduces the amount of training data required and chances of overfitting to a particular size or amplitude input. The scale of extracted features in CNNs depends on the tensor dimensions and convolutional hyperparameters used in each layer of the network [74], [75]. As the input tensor is passed through the network, features are extracted at different scales. Downsampling, such as max-pooling or strided convolutions, reduces the size and spatial resolution of the feature map. As a result, finer resolution features that are present in earlier network layers may be neglected in subsequent coarser resolution layers. This phenomenon is known as the ''semantic gap,'' where the features in different layers of the network represent different levels of abstraction, and finer features may be ignored as the network learns more complex representations [74], [75].

This semantic gap is of extreme importance in perception models because features of smaller scale can be overlooked in deeper convolutions. For example, in computer vision, features from distance or small objects may be lost as the input image is downsampled throughout multiple convolutions. This means the model loses the ability to identify a class at various sizes and distances and therefore is quite limited in its uses.

This same principle applies to SSL and may be even more important. This is because SSL algorithms must differentiate between direct and indirect signals (such as reflections, reverberations, and diffractions). These indirect signals generally have a similar (or identical) wave pattern as their source's direct signals, but at a reduced amplitude and with a phase shift. From a neural network's perspective, features of quieter or indirect signals are the equivalent to another source that is further away; the distinguishing patterns are the same but of different amplitude and phase. To differentiate between direct and indirect signals, a SSL model should: 1) identify all signals from the same source; and 2) isolate the direct signal based on its scale relative to the indirect signals. For both these steps, the model must understand feature scaling.

To address the semantic gap, specific architectures have been proposed, such as U-Net and Feature Aggregators. Various studies have conducted performance comparisons between Feature Aggregation and U-Net architectures, demonstrating that feature aggregation networks achieve high performance in various image segmentation tasks [76]. Although, tasks such as brain tumor segmentation, which have more emphasis on fine-grained details, benefit from using a U-Net which has been trained on a sufficiently large dataset [77], [78]. Additionally, feature aggregation networks exhibited enhanced computational efficiency, reduced memory footprint, and the ability to achieve high accuracy with smaller training datasets. The latter was demonstrated in a study that examined the effectiveness of feature aggregation networks in the context of image segmentation [79]. The demand for less data is particularly crucial in the context of sound detection models as each class of signal must be sampled at many angles or locations; with variables like room size/shape, wall material, and objects in the room affecting signal reflections and reverberations.

Therefore, the choice between feature aggregation and encoder architectures ultimately depends on the desired tradeoff between efficiency and accuracy. Some tasks may require more emphasis on fine-grained details and thus benefit from the use of U-Net, while others may prioritize computational efficiency and simplicity, making feature aggregation a more suitable option [78]. This study focuses on feature aggregation due to its practicality in real world applications. Large computational cost, memory footprint, and training dataset requirements make sound detection U-Nets impractical for sound detection.

B. FEATURE AGGREGATION

The purpose of feature aggregation is to combine features from various convolutions throughout the network to improve scaling, overfitting, and exploding or vanishing gradients [80]. Feature aggregation has three sequential steps: resampling inputs to match shapes, aggregate inputs (weighted averaging or concatenation), and convolution of the aggregated tensor [80]. These processes are completed

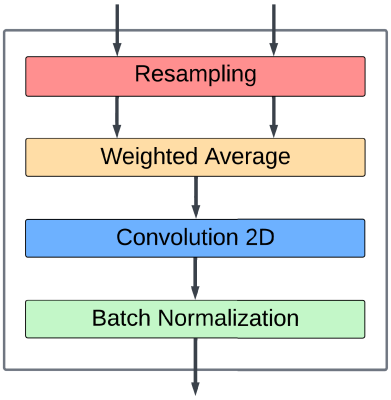


FIGURE 1. Aggregation node [80] diagram illustrating sequential procedures performed within each feature aggregation node.

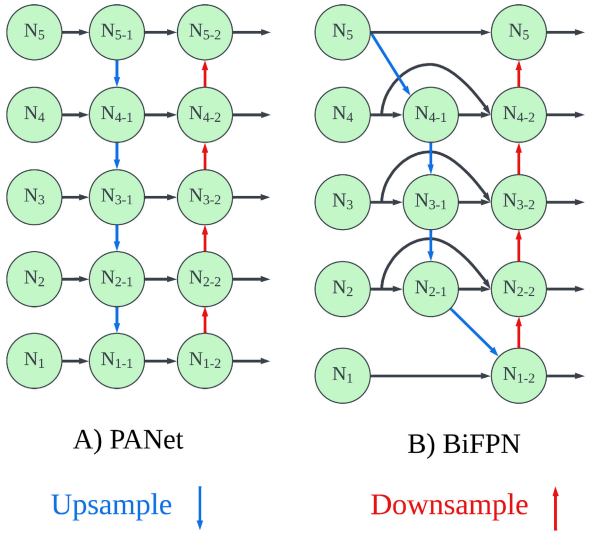


FIGURE 2. Example diagrams of PANet [89] and BiFPN [80] feature aggregators with five scales, where N_i is a feature level with resolution $1/2^i$ of the input image, and N_{i-1} and N_{i-2} are intermediate and output features at level i, respectively. PANet involves propagation of high-level features from the top-down to the bottom-down pathway. BiFPN additionally introduces bidirectional cross-scale connections and weighted feature fusion.

inside a TensorFlow sub-model called a "Node", as is illustrated in Fig. 1. Multiple nodes can be connected residually within an aggregator, allowing this process to be repeatedly executed throughout an elaborate structure.

Resampling within the aggregator node consists of two processes, downsampling and upsampling [80]. Downsampling reduces the resolution of feature maps, resulting in a smaller spatial dimension but a larger number of channels. It is typically done through operations such as max pooling or strided convolutions. In contrast, upsampling increases the resolution of feature maps by interpolating the values and can be achieved through techniques such as transposed convolution (aka deconvolution) or interpolation [80], [81], [82]. One commonly used method for upsampling tensors in computer vision and image processing is bilinear interpolation.

While transposed convolution offers advantages, such as the potential for improved performance through learned parameters, it is known to cause checkerboard artifacts due to non-unit strides [83], [84], [85]. These artifacts can be avoided by using bilinear interpolation, which is also more computationally efficient than transposed convolution. Furthermore, bilinear interpolation is preferred over other interpolation methods like nearest-neighbor interpolation because it produces smoother and more natural-looking results [86]. Nearest-neighbor interpolation simply selects the nearest pixel value without considering its neighboring pixels, which can result in jagged edges and other artifacts [81]. This spatial preservation of features is vital for the resampling of both images and spectrograms [7], [10]. Additionally, bilinear interpolation can be easily extended to higher dimensions, such as 3D volumes or tensors [79]. It is relatively easy to implement and understand, making it a popular choice for spatial features [76]. This study utilizes bilinear interpolation over other upsampling methods due to the reasons listed above.

Industrial computer vision object detection models that demonstrate resampling are YOLO (You Only Look Once) [81] and SSD (Single Shot MultiBox Detector) [82]. In YOLO, downsampling is performed using strided convolutions, while upsampling is achieved using transposed convolution. In SSD, downsampling is performed using max pooling, and upsampling is done using deconvolution layers. Comparatively, this study performs downsampling using strided convolutions and upsampling through bilinear interpolation. Our choice for resampling methods is based on [80]; a study that examines the efficiency of aggregation methods in computer vision and uses bilinear interpolation during its more efficient aggregation. The difference in resampling approaches demonstrates that there is not one universal method for the resampling process. The choice of algorithm for aggregation steps can vary depending on desired computational cost and desired performance.

Once resampling is complete, the tensors are aggregated using weighted averaging. This is more efficient than the conventional method of concatenation as it results in a smaller tensor [87]. Weighted averaging allows the model to directly contrast features derived at distinct scales and processing depths, diversifying the scale and refinement level of features used for final predictions [88]. The weights are trainable variables, allowing the model to optimize the aggregation.

As seen in Fig. 2, feature aggregators connect nodes via residual connections, creating a complex residual network. A benefit of these residual connections is the minimization of vanishing/exploding gradients throughout the network [88]. Residual connections enable the gradient to flow through the network more easily, stabilizing the training process. In Path Aggregation Network (PANet), residual connections propagate high-level semantic features from the top-down pathway to the bottom-up pathway, enabling the network to generate highly detailed object proposals at multiple scales [88], [89].

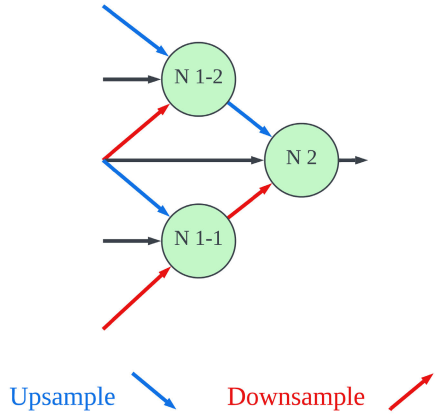


FIGURE 3. Example diagram of SEN feature aggregator encoding five scales down to one scale.

Feature aggregator architectures are classified based on their number of nodes and connection order, which varies with the feature extraction depth and structure. However, aggregation of additional feature scales results in computational overhead, necessitating a tradeoff between higher aggregation and prediction speed [80], [88]. This is why recent object detection models transitioned from Feature Pyramid Network (FPN) to PANet for enhanced performance [76], yet subsequent developments have focused on more compact alternatives such as Neural Architecture Search Feature Pyramid Network (NAS-FPN) [90] and Weighted Bi-Directional Feature Pyramid Network [80].

C. SCALE ENCODER NETWORK

The novel aggregator introduced in this study is Scale Encoder Network (SEN). The goal of SEN is to reduce aggregation computational complexity from the node aggregation process while still enabling the neural network to weigh scales in a residual manner. Its premise is based on the idea that encoding multiple scales into one scale results in less nodes but still addresses the semantic gap.

Aggregators such as FPN, PANet, and BiFPN essentially update a constant number of scales throughout their process. If the backbone of the network has N nodes (or scales), the final layer of these aggregators also has N nodes and outputs. SEN, on the other hand, compresses multiple scales into one, as seen in Fig. 3. In SEN, consecutive aggregation layers reduce N until it is equal to the desired number of outputs. In Fig. 3, five initial scales are compressed into two nodes, then one node during aggregation.

In [80], seven backbone outputs feed into aggregators. In this case, PANet adds 14 resamplings, weighted averages, and convolutions to a network of only seven convolutions. A SEN with a compression width of two would have a first layer of three nodes and a second layer of one node. In models like DarkNet [91], there can be over 50 convolutional blocks in the backbone, and a SEN aggregator will have a huge impact on the computational cost of feature aggregation.

The number of scales between nodes in a SEN layer is referred to as the compression width W. For the SEN nodes in Fig. 3, each node is thought to be a feature scale that is at, one above, and one below of the scales of its inputs. Thus, the compression width W of this encoding at each node can be thought of as 2, since the max difference between the input scale levels is 2. There are several factors to consider when choosing it. The first factor is the volume and scope of resampling. While downsampling results in data loss, upsampling necessitates data approximation. In many feature aggregators (PANet, FPN, BiFPN, NAS-FPN), scales are never resampled more than one scale at a time [80], [90]. Presumably, this is to minimize data loss and approximation throughout the network. In an attempt to not resample one scale too much, we chose to use the middle backbone scale for SEN outputs. Connection overlaps should also be considered; a compression width of one can result in many overlapping connections, which could lead to repetitive calculations and a preference for particular scales. To compare the effects of various compression width sizes, this study evaluates two different SEN designs; which are explained further in Section III.

D. OBJECT DETECTION ARCHITECTURE

Multi-Task Learning (MTL) refers to the process of training neural networks to make predictions for multiple tasks, such as object localization and classification, simultaneously [92]. This enables models to train variables while also cross-referencing losses from each task. By collectively downplaying noise and anomalous patterns that one task may have overemphasized, each task contributes evidence for the applicability of features. This prevents overfitting by focusing on crucial traits. An additional benefit to MTL is eavesdropping, which occurs when information obtained from simple tasks is used to finish a complex task [92]. For instance, filters used in image segmentation to identify an object's class provide information about the object's shape, which can be used to calculate the object's coordinates.

Usually, object detection models consist of the three steps of feature extraction, feature aggregation, and prediction [81], [91]. Collectively, these processes extract relevant features from input tensors, combine those features into a single representation, and then use that representation to predict the presence and location of objects within each tensor's unique coordinate system. By following these three steps, object detection neural networks can achieve cutting-edge performance for a variety of object detection applications.

The input tensors are analyzed in the feature extraction stage (or "the backbone") to identify relevant data patterns, typically using convolutional layers. A hierarchy of increasingly complex patterns is produced as the backbone processes the tensor [93], [94]. High-level estimator performance is still not assured, even though these extracted features are a more useful representation for predictions than raw data. In the feature aggregation stage, the object detection model will produce feature maps that are resistant to changes in scale, translation, and rotation [80], [88], [94]. The feature

aggregation stage is covered in detail in Section II-B. During the last stage, prediction (or ''the neck''), aggregated features are transformed into an entire set of predictions. In computer vision's object detection models, separate dense branches frequently perform classification and box coordinate regression for the detected objects [81], [82].

For computing final predictions in object detection, anchors, like those found in YOLO and SSD, have become standard [81], [88]. In order to localize objects in an image, anchors are a collection of predefined bounding boxes with various scales and aspect ratios. Anchors streamline the prediction process by splitting the task into two distinct tasks; determining whether an anchor box contains an object or not, and adjusting the anchor box coordinates to fit any present objects. This method enables the network to generalize objects of specific shapes and sizes while reducing the number of trainable parameters, speeding up parameter optimization. The network only predicts the presence and location of objects in a fixed set of boxes rather than predicting the precise location of each object across the entire image, making anchors computationally efficient. Anchors do not exist yet for SSL and sound detection models. However, feature aggregation is essential to proper function of anchors. Incorporating feature aggregation into sound detection architectures allows for the use of sound signal anchors in succeeding research, that is expected to substantially improve the localization and classification of sound sources at various amplitudes and phases [81], [82], [91].

This study uses a baseline model consisting of only a basic backbone and head as a control model. The backbone of this model consists of three sequential convolutions and the head has two branches, each with two sequential dense layers. As will be elaborated upon in the next section, Methodology, various feature aggregators are inserted between the backbone and neck of the model; replicating the computer vision object detection architecture described in this section.

III. METHODOLOGY

To isolate the effects of feature aggregation on SSL models, this section will introduce four new SSL models that integrate various aggregators into a baseline model (taken from [7]) and then trained with identical datasets [95] and preprocessing [7]. The baseline architecture SELDnet can be seen in Fig. 4 and the proposed models with aggregation in Fig. 5. Importantly, SELDnet is only used to demonstrate the proof of concept. It was selected due to its publicly available source code and training datasets, which ensure reproducibility and verification of the results. However, the created feature aggregation submodels can be integrated with more recent Sound Event Localization and Detection networks (e.g., SELD3DNET [29]), as explained in our publicly available Feature Aggregator Library [96].

A. DEVELOPMENT

This study used Keras library with TensorFlow's Functional API backend to implement and test the feature aggregation

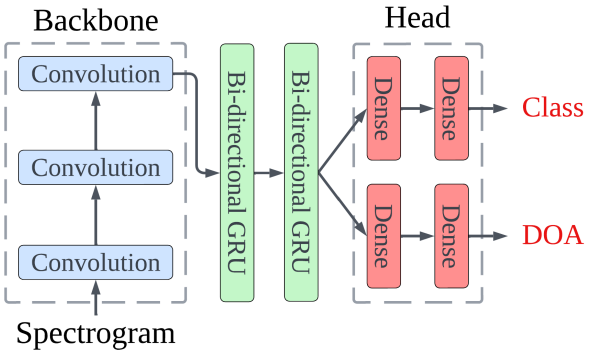


FIGURE 4. Illustration of the baseline SELDnet [7].

models. This API was chosen because of its flexibility for creating non-sequential neural networks. Feature aggregation nodes are Model class objects that incorporate sequential resampling, weighted averaging, and convolutional layers for processing input tensors. As the current version of the TensorFlow Functional API does not include a weighted averaging layer, we utilized a custom layer of the Layer class, with the weights designated as trainable variables. Subsequently, multiple node sub-models were interconnected to create Feature Aggregator sub-models, which were then integrated into the main model. This design allows nodes to be easily arranged to quickly create aggregators, and aggregators to be efficiently integrated into larger architectures.

Each model was trained for a max of 1000 epochs with early stoppage if the SELD score (see Section IV-C) on the test split does not improve for 100 epochs. This early stoppage is to prevent network over-fitting. For training loss, we utilized a weighted combination of binary cross-entropy for classification and MSE for localization with an Adam optimizer with default parameters [97].

Simple 1×1 convolutions were used in the feature aggregation nodes due to their ability to reduce dimensionality of feature maps [98] and apply nonlinear transformations [99]. In terms of dimensionality reduction, the convolutional operation with a single filter and a stride of 1 in the time axis can be used to generate a new feature map with a reduced number of channels, which is particularly beneficial in deep neural networks where the number of feature maps can quickly become large and computationally expensive to process [14]. As for nonlinear transformation, the convolutional operation with multiple filters and a nonlinear activation function, such as ReLu or softmax, are applied to feature maps to increase their expressive power and improve the performance of the neural network [99].

B. NETWORK ARCHITECTURES

SELDnet is a MTL CRNN without feature aggregation which serves as a baseline architecture for this work. It simultaneously predicts the presence of multiple classes and their relative positions in 3D Cartesian coordinates. It has been chosen for a few reasons. First, this is currently

a state-of-the-art architecture that performed well in a distinguished study by [7]. Second, the architecture design allows for easy integration of feature aggregators compared to non-sequential networks. Third, the hyperparameters have already been tuned, allowing this study to focus on tuning feature aggregators.

The feature extraction in SELDnet is done by three sequential convolutions layers while the prediction stages are completed by two branches of two dense layers, respectively. Fig. 5 presents the proposed architectures by this study. Figs. 5 (a)-(b) display SELDnet with the established PANet and BiFPN aggregators. Figs. 5 (c)-(d) illustrate SELDnet with two variations of SEN. The first variation, Fig. 5 (c), incorporates an aggregator with two SEN layers with compression width of one. The second variation, Fig. 5 (d), demonstrates the SELDnet with a single SEN layer of compression width two.

The feature aggregators to implement onto SELDnet chosen for this study are PANet, BiFPN, and SEN. PANet and BiFPN were selected because of their well-established track record in popular object detection models, including YOLO and SSD [80], [82]. SEN is a new aggregator developed by this study. To evaluate the effects of the compression width value, two models with SEN are tested: one test model with a compression width of one ($SEN_{W=1}$) and the other with a compression width of two ($SEN_{W=2}$). The SEN model with a compression width of one uses two intermediate scales and these scales sizes are the averaged dimensions of their input tensors. All of these aggregators vary in number of nodes and connection patterns, allowing for analysis and speculation of optimal approaches for feature aggregation design.

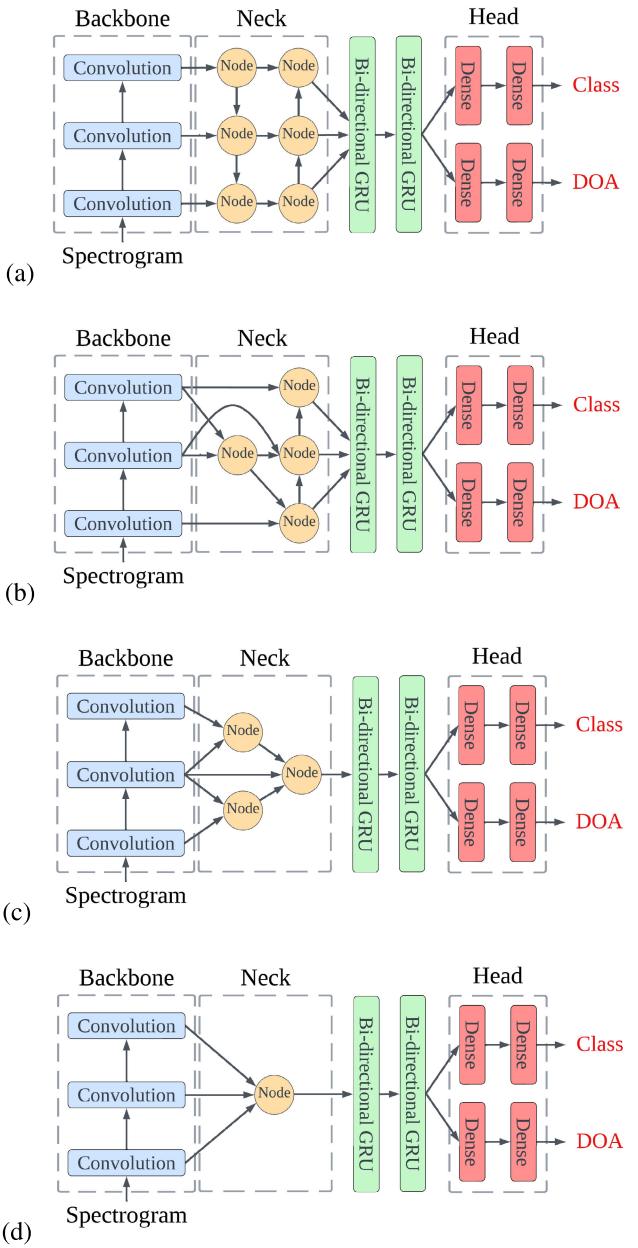


FIGURE 5. Diagrams of final model architectures proposed by this study. Subfigures a, b, c, and d illustrate SELDnet with PANet, BiFPN, $SEN_{W=1}$, and $SEN_{W=2}$, respectively.

IV. EVALUATION

A. DATASET

The REAL dataset, compiled by [95], serves as a valuable resource for research on sound event detection (SED) and localization [7]. The dataset consists of 216 uncompressed WAV multichannel audio recordings, each lasting 30 seconds, captured in a university corridor surrounded by classrooms during working hours. The settings represent common real-life acoustic environments and feature diverse overlapping sound sources and backgrounds.

More specifically, the data comprises impulse responses (IRs) from the environment, measured using an Eigenmike spherical microphone array. The IRs were generated by slowly moving two loudspeakers (continuously playing a maximum length sequence) in circular trajectories around the array, one elevation at a time. These trajectories had radii of 1 meter (m) and 2 m from the array. At a distance of 1 m, the elevation ranged from $-40°$ to $40°$ in $10°$ increments, and at 2 m, the elevation varied from $-20°$ to $20°$ with the same increments. The dataset contains spatial coordinates for the loudspeakers and microphones, along with annotations that provide information about the temporal boundaries, classification, and spatial coordinates of sound events present in each recording. It encompasses 8 distinct sound categories, including car horn, dog bark, drilling, engine idling, gun shot, jackhammer, siren, and street music.

B. PREPROCESSING

In order to discern the specific impact of feature aggregation within our framework, the data preprocessing methodology employed is identical to the baseline model study [7] and uses the code from this reference repository. The preprocessing stage involves the following steps:

1) Raw audio files were de-noised using band-pass filtration to remove low and high-frequency noise. This

method is effective against typical noise in recording environments [7]. Signals were then downsampled to 16 kHz, reducing computational complexity and ensuring efficient data preprocessing for future stages.

2) The STFT of the preprocessed audio signals were computed to create a detailed time-frequency representation [11]. A window size of 1024 samples and hop size of 256 samples provided the optimal temporal and spectral details in the resulting spectrogram [100]. The magnitude and phase from the STFT spectrogram were used as separate input features to the neural network to improve the localization performance. Extracting these features from the raw waveforms requires minimal pre-processing of the data.
3) To bolster the training dataset and improve model generalization, data augmentation techniques were employed, including random time shifting, frequency shifting, and amplitude scaling [79]. This enriched dataset enhanced the model's adaptability and performance in varied acoustic scenarios.

C. METRICS

Each model's performance is evaluated using several metrics that measure the accuracy of the SED and the sound event DOA estimation. The SED metrics are F-score and Error Rate. F-score (F) is a widely used metric for binary classification problems that measures the balance between precision and recall [101]. It is defined as the harmonic mean of precision and recall. In the context of SED, True Positives (TP) refer to the correctly detected events, False Positives (FP) refer to the events that were incorrectly detected, and False Negatives (FN) refer to the events that were missed by the model [7]. F-score is a useful metric because it considers both the number of correctly detected events and the number of missed and false alarms. A higher F-score indicates a better performance. The F-score for predicting the presence of classes in each k one-second segment is defined as

$$F = \frac{2 \cdot \sum_{k=1}^{K} TP(k)}{2 \cdot \sum_{k=1}^{K} TP(k) + \sum_{k=1}^{K} FP(k) + \sum_{k=1}^{K} FN(k)} \quad (1)$$

where, for each one-second segment k, $TP(k)$ is the number of true positives; $FP(k)$ is the number of false positives; and $FN(k)$ is the number of false negatives. True positives are correctly predicted sound event classes which are present in the segment. In contrast, false positives are sound event classes erroneously predicted to be within the segment and false negatives are sound event classes present in the segment but failed to be detected.

Error Rate (ER) is another commonly used metric for SED, which measures the percentage of incorrectly detected events [7], [102]. ER is calculated as

$$ER = \frac{\sum_{k=1}^{K} S(k) + \sum_{k=1}^{K} D(k) + \sum_{k=1}^{K} I(k)}{\sum_{k=1}^{K} N(k)} \quad (2)$$

where, for each one-second segment k, $N(k)$ is the total number of active sound event classes in the ground truth. $S(k)$, substitution, is the number of times an event was detected at the wrong level and is calculated by merging false negatives and false positives without individually correlating which false positive substitutes which false negative. The remaining false positives and false negatives, if any, are counted as insertions $I(k)$ and deletions $D(k)$, respectively. These values are calculated as follows:

$$S(k) = min(FN(k), FP(k)) \quad (3)$$
$$D(k) = max(0, FN(k) - FP(k)) \quad (4)$$
$$I(k) = max(0, FP(k) - FN(k)) \quad (5)$$

The DOA metrics are DOA error and Frame Recall. DOA error measures the difference between the estimated DOA and the ground truth DOA in degrees for the entire dataset with total number of DOA estimates, D [7]. A lower DOA error indicates a better performance. The error is defined as

$$\text{DOA Error} = \frac{1}{D} \sum_{d=1}^{D} \sigma((x_G^d, y_G^d, z_G^d), (x_E^d, y_E^d, z_E^d)) \quad (6)$$

where (x_E, y_E, z_E) is the predicted DOA estimate, (x_G, y_G, z_G) is the ground truth DOA, and σ is the angle between (x_E, y_E, z_E) and (x_G, y_G, z_G) at the origin for the d-th estimate:

$$\sigma = 2 \cdot \arcsin\left(\frac{\sqrt{\Delta x^2 + \Delta y^2 + \Delta z^2}}{2}\right) \cdot \frac{180}{\pi} \quad (7)$$

with $\Delta x = x_G - x_E$, $\Delta y = y_G - y_E$, and $\Delta z = z_G - z_E$.

Frame Recall (FR), as defined by [7], is a metric used to measure the accuracy of a model's predictions in the context of time frames or segments. It measures the percentage of time frames where the number of estimated and ground truth DOAs are unequal. A higher Frame Recall indicates a better performance and is calculated as

$$FR = \frac{\sum_{k=1}^{K} TP(k)}{\sum_{k=1}^{K} TP(k) + FN(k)} \cdot 100 \quad (8)$$

We used a combined localization and classification score, SELD, to perform early training stoppage. If the SELD score did not improve over 100 epochs, training was terminated to prevent overfitting. SED and DOA scores represent the overall performance of an estimator for sound event detection and localization, respectively. SELD is the average of these scores and functions as a single overarching metric to compare models. A lower value indicates better performance for DOA, SED, and SELD scores. Eqs. (9), (10) and (11) define these metrics [7]:

$$\text{DOA score} = \frac{((\text{DOA Error})/180 + (1 - FR/100))}{2} \quad (9)$$
$$\text{SED score} = \frac{(ER + (1 - F/100))}{2} \quad (10)$$
$$\text{SELD} = \frac{((\text{SED score}) + (\text{DOA score}))}{2} \quad (11)$$

D. BASELINE METHODS

SELDnet is a deep learning architecture developed specifically for combined SED and DOA estimation tasks [7]. SELDnet processes spectral features dedicated to the SED task in parallel with spatial features for the DOA estimation. By leveraging the fusion of convolutional and recurrent layers, SELDnet effectively pinpoints both the occurrence and the spatial origin of a sound event. MSEDnet, derived from SELDnet, is designed for monaural (single-channel) SED. By focusing on the SED task, MSEDnet provides an optimal solution for applications where the sole requirement is event detection without the need for spatial localization [103]. Its architecture is tailored to single-channel audio contexts, ensuring precise event detection. SEDnet stands as a dedicated solution for SED, capitalizing on deep learning techniques [103]. With its architecture centered around event identification, SEDnet excels in scenarios where temporal detection of sound events is the primary objective. It delivers accuracy and efficiency in sound event classification without incorporating spatial estimation components. DOAnet offers a specialized approach towards the spatial dimension of audio signals, focusing solely on DOA estimation [104]. MUSIC (Multiple Signal Classification) is a robust algorithm for DOA estimation. Relying on subspace methods, MUSIC differentiates the signal space from the noise space, facilitating precise DOA predictions for multiple sound sources [105]. Its mathematical foundation and proven efficiency in array signal processing render it as a reliable choice for DOA estimation tasks, even in contexts dominated by neural network models [7].

It is important to note that the comparisons made with these baseline models are not intended to show that integrating the proposed feature aggregators with SELDnet make it outperform all existing SELD networks published to date. The purpose is rather to simply evaluate the quantitative improvement resulted from integrating the proposed feature aggregator with *a Sound Event Localization and Detection network* in the joint classification and localization task. However, the created feature aggregation submodels (available in our publicly available Feature Aggregator Library [96]) can be integrated with other Sound Event Localization and Detection networks, such as SELD3DNET [29], and may show improvement with respect to the model without feature aggregators. Additionally, the study considers single-task methods specialized in SED (MSEDnet and SEDnet) and SSL (DOAnet and MUSIC) to determine if incorporating the feature aggregators can sufficiently boost the performance of SELDnet to make it comparable to these specialized methods in individual SED and SSL tasks.

V. RESULTS

The results in Table 1 indicate that feature aggregation enhanced the model's capacity in both localization and classification of the sound sources. None of the models evaluated in this study outperformed the algorithms specialized for only classification or localization, but all demonstrated a clear improvement in both tasks compared to the original SELDnet.

A. CLASSIFICATION

For the SELDnet variants, classification improvements were clear but minimal. All of the *ER* scores are closely clustered, making it difficult to determine the extent to which different feature aggregation designs affected scores. However, the F and SED scores imply that aggregation did improve SELDnet's ability to classify in a manner comparable to MSEDnet and SEDnet. The dataset may impose restrictions on these scores, preventing architectural design from displaying a large impact. Datasets can limit deep learning model performance through factors such as data size, quality, class imbalances, noisy or biased labeling, and distribution mismatches; all of which hinder the model's ability to achieve over a certain score. Compared to MSEDnet, both SELDnet+PANet and SELDnet+SEN$_{W=1}$ performed marginally worse. These two models performed marginally better than SEDnet, whereas SELDnet with BiFPN and SEN$_{W=2}$ performed comparably to SEDnet. With respect to SEDnet, SELDnet with BiFPN and SEN$_{W=2}$ scored slightly better on *ER*, slightly worse on F-Score, but achieved the same overall SED score. Overall, the models with aggregation provide better performance; they are associated with lower *ER*, higher F-Score, and lower SED score.

B. LOCALIZATION

The improvements in localization scores of models with aggregation are notable. Compared to SELDnet, all the networks with aggregation considerably improved performance, as was expected. All aggregated models had considerably lower DOA Errors, higher Frame Recall, and lower DOA score. SELD+PANet performed particularly well in DOA estimation, with metric scores that stand out from the other clustered aggregation DOA scores. This clear DOA estimation boost suggests that feature aggregation enhances the distinguishing of various sound signals, such as reverberations, diffractions, and direct signals. The increased robustness to indirect signals, observed after introducing aggregators into SELDnet, is attributable to the enhanced feature scaling. Indirect signals, such as reflections, can exhibit comparable wave patterns to direct signals at lesser amplitudes. Therefore, one would anticipate that a better comprehension of feature scales would enhance the differentiation between direct and indirect signals.

Compared to DOAnet, SELDnet and its variants have a higher frame recall and overall DOA Error; signifying that they excel at correctly identifying DOAs within individual time frames while maintaining consistency with the ground truth DOAs. This indicates that these models have difficulty minimizing the overall disparity between predicted and ground truth DOA compared to DOAnet, but consistently capture signal patterns with respect to time.

TABLE 1. Classification and Localization Scores of Test Architectures Compared to Control Algorithms. The best-performing metrics are highlighted in bold and the secondary metrics are underlined. Arrows indicate the desired direction for better performance: ↓ for lower values and ↑ for higher values.

Architecture \ Algorithm	Classification			Localization			Joint Task
	ER ↓	F-Score ↑	SED ↓	DOA Error ↓	Frame Recall ↑	DOA Score ↓	SELD Score ↓
SELDnet [7]	0.41	60.5	0.40	26.9	65.3	0.25	0.325
SELDnet + PANet	0.36	65.1	0.35	11.8	**72.5**	**0.17**	**0.262**
SELDnet + BiFPN	0.37	63.0	0.37	15.2	68.6	0.20	0.285
SELDnet + $SEN_{W=1}$	0.36	63.6	0.36	14.3	69.5	0.19	0.277
SELDnet + $SEN_{W=2}$	0.37	62.5	0.37	16.1	68.0	0.20	0.289
MSEDnet [103]	**0.35**	**66.2**	**0.34**	-	-	-	-
SEDnet [103]	0.38	64.6	0.37	-	-	-	-
DOAnet [104]	-	-	-	**6.30**	46.5	0.29	-
MUSIC [105]	-	-	-	36.3	-	-	-

It is likely that DOAnet's inconsistency with respect to time is a result of an inability to distinguish direct signals from reflections, reverberations, and diffractions. The data implies that SELDnet variants (developed by this study) are adept at pinpointing the precise time instances when sound sources appear, leading to an improved ability to distinguish between multiple signals. Furthermore, by consistently capturing signal patterns over time, these models are likely to be more robust in dynamic soundscapes where the number of sound sources and noise interference can vary. This trait is vital in real-world applications where sound sources often overlap and vary in number and characteristics.

C. JOINT CLASSIFICATION AND LOCALIZATION

The SELD scores indicate that, regardless of aggregator design, feature aggregation improves the function of joint sound classification and localization models. Although aggregators with more nodes outperformed the single node SEN model, this modest aggregator demonstrated that even minimal aggregation can counter-act the negative effects of the semantic gap. Clearly, PANet's in-depth and equal processing of all scales is optimal for performance. However, as previously discussed, this approach can be computationally demanding, which may prove adverse for certain situations.

D. AGGREGATOR COMPARISON

This section will compare aggregators using two metrics: their overall percentage improvement on SED, DOA, and SELD scores and the percentage improvement per node. The percentage improvement per node is intended as a metric to quantify the efficiency of aggregator designs. As can be seen in Table 2, although some aggregators have better overall improvement, others have a better improvement ratio per node, implying a more efficient connection design.

The obvious outlier is $SEN_{W=2}$. The results for this model imply that any aggregation helps counteract the semantic gap. The collective results indicate that less nodes result in a higher percentage improvement per node. However, in this particular $SEN_{W=2}$ aggregator, the magnitude of improvement per feature aggregator node must be taken with a grain of salt due to the simplicity of the aggregator. SELDnet is an unusually compact neural network, and most real-world models are much deeper (such as Darknet 53 with a backbone of 53 convolutional layers). For backbones deeper than three layers, which is the case for most models, SEN aggregators with a compression width of two would involve more than one node. We hypothesize that the use of a single node causes this $SEN_{W=2}$ to seem disproportionally effective per node because the overall score change is divided by one. The overall improvement from this $SEN_{W=2}$ aggregator is a testament to the effects of having any aggregator, but the results for improvement per node are skewed due to division by one. Nevertheless it is interesting that compared to BiFPN, this single node performed comparatively in overall SED and DOA percentage improvement and slightly worse in overall SELD. It is important to note that the DOA and SED scores of BiFPN and $SEN_{W=2}$ are the same because of rounding, but the actual difference is seen in the SELD score. As will be discussed later in this section, we attribute this similar performance to the efficacy of encoder aggregators of SEN.

After removing this outlier and comparing PANet, BiFPN and $SEN_{W=1}$, the next clear take away is PANet's overall percentage improvement. PANet was not as efficient as $SEN_{W=1}$, but the overall improvement is substantial compared to all other aggregators. This is attributable to the in-depth processing at every scale, which is the most comprehensive design for addressing the semantic gap. The lower efficiency per node is likely the result of certain scales not requiring as much processing as is actively occurring.

$SEN_{W=1}$ is clearly an efficient design, with the second highest overall percentage improvements and highest percentage improvement per node, when excluding the outlier $SEN_{W=2}$. This efficiency per node is attributable to the efficacy of the encoder approach. This approach allows for even weighing and consideration of all scales, like PANet, without leading to uneven processing of each scale, as seen in BiFPN. This equal evaluation of all scales with reduced nodes leads to an efficient aggregation process. BiFPN's results indicate that the design is not the best performing or most efficient. Architectures with more trainable parameters, such

TABLE 2. Comparison of Aggregator Effects on Baseline SELDnet Scores. The best-performing metrics are highlighted in bold.

Aggregator	# Nodes	SED % Improvement		DOA % Improvement		SELD % Improvement	
		Overall	Per Node	Overall	Per Node	Overall	Per Node
PANet	6	**12.5**	2.08	**32.0**	5.33	**19.4**	3.23
BiFPN	4	7.50	1.88	20.0	5.00	12.3	3.08
$SEN_{W=1}$	3	10.0	3.3	24.0	8.00	14.8	4.92
$SEN_{W=2}$	1	7.50	**7.50**	20.0	**20.0**	11.1	**11.1**

as the additional nodes in PANet, can be trained to outperform BiFPN on the test dataset. It is interesting that the SEN models performed well in comparison with BiFPN, which has more nodes than both SEN models. We hypothesize that BiFPN overemphasized one scale due to the scale's extra nodes, whereas SEN created a compressed representation with equal weighting of all scales.

VI. CONCLUSION

The results indicate that, regardless of the aggregator's design, feature aggregation can significantly improve the performance of neural networks in SELD and specially SSL. Compared to SELDnet, all networks with feature aggregation showed considerable improvement in localization, classification, and the joint task of SELD. The addition of feature aggregators particularly enhanced the localization task, resulting in significantly higher frame recall and notably lower DOA score compared to single-task networks such as DOAnet, which specializes in DOA prediction. The performance in classification was also competitive with specialized networks such as MSEDnet and SEDnet.

A balance must be struck between computational expense and performance when deciding on an aggregator. While examining the available aggregators, SEN and PANet stand out as the most cost-effective and robust, respectively. The difference in aggregator performances indicates that when performing feature aggregation, it is best to equally emphasize all scales. Future research may delve deeper into an assortment of topics, such as the establishment of anchors for spectrograms and the development of more complex SEN designs (such as stacking SEN after FPN or PANet). The created feature aggregation submodels can be additionally integrated with more recent Sound Event Localization and Detection models to compare performance.

ACKNOWLEDGMENT

The authors would like to express their gratitude to Tampere University of Technology and its licensors for granting permission to use the code for the sound event localization and detection using convolutional recurrent neural network method/architecture, which is available in the GitHub repository with the handle "seld-net" found at https://github.com/sharathadavanne/seld-net. This code was described in the article titled "Sound Event Localization and Detection of Overlapping Sources Using Convolutional Recurrent Neural Network" [7].

They also acknowledge and honor the non-commercial nature of this grant and affirm their commitment to preserving the copyright notice in all reproductions of this work. Furthermore, they are grateful to the original source of the work, the Audio Research Group, Laboratory of Signal Processing, Tampere University of Technology. Availability of data, material, or code: https://gitlab.com/dsim-lab/paper-codes/feature-aggregation-for-neural-networks.

REFERENCES

[1] S. Chakrabarty and E. A. P. Habets, "Broadband DOA estimation using convolutional neural networks trained with noise signals," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust. (WASPAA)*, 2017, pp. 136–140, doi: 10.1109/WASPAA.2017.8170010.

[2] S. E. Chazan, H. Hammer, G. Hazan, J. Goldberger, and S. Gannot, "Multi-microphone speaker separation based on deep DOA estimation," in *Proc. 27th Eur. Signal Process. Conf. (EUSIPCO)*, A Coruna, Spain, Sep. 2019, pp. 1–5.

[3] A. Xenaki, J. B. Boldt, and M. G. Christensen, "Sound source localization and speech enhancement with sparse Bayesian learning beamforming," *J. Acoust. Soc. Amer.*, vol. 143, no. 6, pp. 3912–3921, Jun. 2018.

[4] X. Li, L. Girin, F. Badeig, and R. Horaud, "Reverberant sound localization with a robot head based on direct-path relative transfer function," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2016, pp. 2819–2826.

[5] P. Chiariotti, M. Martarelli, and P. Castellini, "Acoustic beamforming for noise source localization—Reviews, methodology and applications," *Mech. Syst. Signal Process.*, vol. 120, pp. 422–448, Apr. 2019.

[6] P.-A. Grumiaux, S. Kitić, L. Girin, and A. Guérin, "A survey of sound source localization with deep learning methods," *J. Acoust. Soc. Amer.*, vol. 152, no. 1, pp. 107–151, Jul. 2022.

[7] S. Adavanne, A. Politis, J. Nikunen, and T. Virtanen, "Sound event localization and detection of overlapping sources using convolutional recurrent neural networks," *IEEE J. Sel. Topics Signal Process.*, vol. 13, no. 1, pp. 34–48, Mar. 2019.

[8] S. Rasheed, S. Ahmad, M. A. Khan, and M. Shahid, "Sound source localization using support vector machine," *Appl. Acoust.*, vol. 74, no. 5, pp. 635–643, 2013.

[9] Y. Jiang, X. Zhang, and D. Zhang, "A novel sound source localization method based on radial basis function neural network," *Appl. Acoust.*, vol. 85, pp. 91–98, Jan. 2014.

[10] S. Adavanne, A. Politis, and T. Virtanen, "Deep learning based direction of arrival estimation for music and speech," in *Proc. Int. Conf. Digital Audio Effects (DAFx)*, 2018.

[11] D. Xie, X. Jiang, Y. Wang, and Y. Xu, "DOA estimation for acoustic source arrays based on convolutional recurrent neural networks," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 28, pp. 236–248, 2019.

[12] S. S. Kushwaha, I. R. Roman, M. Fuentes, and J. P. Bello, "Sound source distance estimation in diverse and dynamic acoustic conditions," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust. (WASPAA)*, Oct. 2023, pp. 1–5.

[13] Y. Zhang, W. Wang, Z. Wu, and Y. Li, "Residual convolutional neural network for sound source localization," *Appl. Acoust.*, vol. 176, Jan. 2020.

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

[15] Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative Adversarial Nets," in *Proc. Adv. Neural Inform. Process. Syst. (NIPS)*, 2014.

[16] R. Ranjan, S. Jayabalan, T. N. T. Nguyen, and W. S. Gan, "Sound event detection and direction of arrival estimation using residual net and recurrent neural networks," in *Proc. Detection Classification Acoustic Scenes Events Workshop (DCASE)*, 2019.

[17] S. Lee, H. Yang, H. Choi, and W. Seong, "Zero-shot single-microphone sound classification and localization in a building via the synthesis of unseen features," *IEEE Trans. Multimedia*, vol. 24, pp. 2339–2351, 2022.

[18] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," 2014, *arXiv:1409.0473*.

[19] J. Kim and M. Hahn, "Voice activity detection using an adaptive context attention model," *IEEE Signal Process. Lett.*, vol. 25, no. 8, pp. 1181–1185, Aug. 2018.

[20] Y. Mao, Y. Zeng, H. Liu, W. Zhu, and Y. Zhou, "ICASSP 2022 L3DAS22 challenge: Ensemble of Resnet-conformers with ambisonics data augmentation for sound event localization and detection," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2022, pp. 9191–9195.

[21] Y. He, S. Shin, A. Cherian, N. Trigoni, and A. Markham, "Sound3DVDet: 3D sound source detection using multiview microphone array and RGB images," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2024, pp. 5496–5507.

[22] D. Fedorishin, D. Dayal Mohan, B. Jawade, S. Setlur, and V. Govindaraju, "Hear the flow: Optical flow-based self-supervised visual sound source localization," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2023, pp. 2277–2286.

[23] T. Hirvonen, "Classification of spatial audio location and content using convolutional neural networks," in *Audio Engineering Society Convention 138*. New York, NY, USA: Audio Engineering Society, 2015.

[24] Y. Cao, Q. Kong, T. Iqbal, F. An, W. Wang, and M. D. Plumbley, "Polyphonic sound event detection and localization using a two-stage strategy," 2019, *arXiv:1905.00268*.

[25] K. Shimada, Y. Koyama, N. Takahashi, S. Takahashi, and Y. Mitsufuji, "Accdoa: Activity-coupled Cartesian direction of arrival representation for sound event localization and detection," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 915–919.

[26] Q. Wang, J. Du, H.-X. Wu, J. Pan, F. Ma, and C.-H. Lee, "A four-stage data augmentation approach to ResNet-conformer based acoustic modeling for sound event localization and detection," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 31, pp. 1251–1264, 2023.

[27] A. Politis, A. Mesaros, S. Adavanne, T. Heittola, and T. Virtanen, "Overview and evaluation of sound event localization and detection in DCASE 2019," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 29, pp. 684–698, 2021, doi: 10.1109/TASLP.2020.3047233.

[28] L. Xue, H. Liu, Y. Zhou, and L. Gan, "Resnet-conformer network using multi-scale channel attention for sound event localization and detection in real scenes," in *Proc. Int. Conf. Wireless Commun. Signal Process. (WCSP)*, Nov. 2023, pp. 25–30.

[29] P. Mei, J. Yang, Q. Zhang, and X. Huang, "A method of sound event localization and detection based on three-dimension convolution," in *Proc. 7th Int. Conf. Image, Vis. Comput. (ICIVC)*, Jul. 2022, pp. 872–878.

[30] Y. Shul and J.-W. Choi, "CST-former: Transformer with channel-spectro-temporal attention for sound event localization and detection," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2024, pp. 8686–8690.

[31] Y. Min, P. Xin, C. Xu, and H. Liu, "Detection and localization of sound events based on principal components analysis," in *Proc. 2nd Int. Conf. Consum. Electron. Comput. Eng. (ICCECE)*, Jan. 2022, pp. 507–511.

[32] J. Hu, Y. Cao, M. Wu, Q. Kong, F. Yang, M. D. Plumbley, and J. Yang, "A track-wise ensemble event independent network for polyphonic sound event localization and detection," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2022, pp. 9196–9200.

[33] 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, 2023.

[34] S. Cheng, J. Du, Q. Wang, Y. Jiang, Z. Nian, S. Niu, C.-H. Lee, Y. Gao, and W. Zhang, "Improving sound event localization and detection with class-dependent sound separation for real-world scenarios," in *Proc. Asia–Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA ASC)*, Oct. 2023, pp. 2068–2073.

[35] Y. Xie, J. Liu, and Y. Hu, "A polyphonic SELD network based on attentive feature fusion and multi-stage training strategy," in *Proc. 4th Int. Seminar Artif. Intell., Netw. Inf. Technol. (AINIT)*, Jun. 2023, pp. 650–654.

[36] H. Zhu and J. Yan, "A deep learning based sound event location and detection algorithm using convolutional recurrent neural network," in *Proc. Int. Conf. Comput., Inf. Telecommun. Syst. (CITS)*, Jul. 2022, pp. 1–6.

[37] Y. Fujita, Y. Bando, K. Imoto, M. Onishi, and K. Yoshii, "DOA-aware audio-visual self-supervised learning for sound event localization and detection," in *Proc. Asia–Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA ASC)*, Oct. 2023, pp. 2061–2067.

[38] X. Jiang, C. Han, Y. A. Li, and N. Mesgarani, "Exploring self-supervised contrastive learning of spatial sound event representation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2024, pp. 1281–1285.

[39] S.-H. Jo, C. Y. Jeong, and M. Kim, "Sound event localization and detection using spatial feature fusion," in *Proc. 13th Int. Conf. Inf. Commun. Technol. Converg. (ICTC)*, Oct. 2022, pp. 1849–1851.

[40] T. N. Tho Nguyen, D. L. Jones, K. N. Watcharasupat, H. Phan, and W.-S. Gan, "SALSA-lite: A fast and effective feature for polyphonic sound event localization and detection with microphone arrays," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2022, pp. 716–720.

[41] Y. Cao, T. Iqbal, Q. Kong, F. An, W. Wang, and M. D. Plumbley, "An improved event-independent network for polyphonic sound event localization and detection," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 885–889.

[42] M. Yasuda, S. Saito, A. Nakayama, and N. Harada, "6DoF SELD: Sound event localization and detection using microphones and motion tracking sensors on self-motioning human," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2024, pp. 1411–1415.

[43] D. Desai and N. Mehendale, "A review on sound source localization systems," *Arch. Comput. Methods Eng.*, vol. 29, no. 7, pp. 4631–4642, Nov. 2022.

[44] C. Rascon and I. Meza, "Localization of sound sources in robotics: A review," *Robot. Auto. Syst.*, vol. 96, pp. 184–210, Oct. 2017.

[45] C. Pang, H. Liu, and X. Li, "Multitask learning of time-frequency CNN for sound source localization," *IEEE Access*, vol. 7, pp. 40725–40737, 2019.

[46] S. Jiang, L. Wu, P. Yuan, Y. Sun, and H. Liu, "Deep and CNN fusion method for binaural sound source localisation," *J. Eng.*, vol. 2020, no. 13, pp. 511–516, Jul. 2020.

[47] Y. Xu, S. Afshar, R. K. Singh, R. Wang, A. van Schaik, and T. J. Hamilton, "A binaural sound localization system using deep convolutional neural networks," in *Proc. IEEE Int. Symp. Circuits Syst. (ISCAS)*, May 2019, pp. 1–5.

[48] H. Liu, P. Yuan, B. Yang, and L. Wu, "Robust interaural time difference estimation based on convolutional neural network," in *Proc. IEEE Int. Conf. Robot. Biomimetics (ROBIO)*, Dec. 2019, pp. 352–357.

[49] N. Ma, T. May, and G. J. Brown, "Exploiting deep neural networks and head movements for robust binaural localization of multiple sources in reverberant environments," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 25, no. 12, pp. 2444–2453, Dec. 2017.

[50] P. Vecchiotti, N. Ma, S. Squartini, and G. J. Brown, "End-to-end binaural sound localisation from the raw waveform," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 451–455.

[51] J. Wang, J. Wang, K. Qian, X. Xie, and J. Kuang, "Binaural sound localization based on deep neural network and affinity propagation clustering in mismatched HRTF condition," *EURASIP J. Audio, Speech, Music Process.*, vol. 2020, no. 1, pp. 1–16, Dec. 2020.

[52] M. J. Bianco, S. Gannot, and P. Gerstoft, "Semi-supervised source localization with deep generative modeling," in *Proc. IEEE 30th Int. Workshop Mach. Learn. Signal Process. (MLSP)*, Sep. 2020, pp. 1–6.

[53] Q. Nguyen, L. Girin, G. Bailly, F. Elisei, and D.-C. Nguyen, "Autonomous sensorimotor learning for sound source localization by a humanoid robot," in *Proc. IROS Workshop Crossmodal Learn. Intell. Robot. Conjunct. With (IEEE/RSJ IROS)*, Oct. 2018, pp. 1–4.

[54] J. Choi and J.-H. Chang, "Convolutional neural network-based direction-of-arrival estimation using stereo microphones for drone," in *Proc. Int. Conf. Electron., Inf., Commun. (ICEIC)*, Jan. 2020, pp. 1–5.

[55] T.-H. Tan, Y.-T. Lin, Y.-L. Chang, and M. Alkhaleefah, "Sound source localization using a convolutional neural network and regression model," *Sensors*, vol. 21, no. 23, p. 8031, Dec. 2021.

[56] Y. Pamungkas and Y. Rais, "Implementation of real-time sound source localization using TMS320C6713 board with interaural time difference method," in *Proc. 2nd Int. Seminar Mach. Learn., Optim., Data Sci. (ISMODE)*, Dec. 2022, pp. 269–274.

[57] Q. Li, X. Zhang, and H. Li, "Online direction of arrival estimation based on deep learning," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 2616–2620.

[58] F. Grondin, J. Glass, I. Sobieraj, and M. D. Plumbley, "Sound event localization and detection using CRNN on pairs of microphones technical report," Massachusetts Inst. Technol. Univ. Surrey, USA, 2019, pp. 1–5.

[59] S. Chakrabarty and E. A. P. Habets, "Broadband DOA estimation using convolutional neural networks trained with noise signals," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust. (WASPAA)*, Oct. 2017, pp. 136–140.

[60] I. M. Velázquez, Y. Ren, Y. Haneda, and H. M. Pérez-Meana, "DOA estimation for spherical microphone array using spherical convolutional neural networks," in *Proc. IEEE 10th Global Conf. Consum. Electron. (GCCE)*, Oct. 2021, pp. 510–511.

[61] M. J. Bianco, S. Gannot, E. Fernandez-Grande, and P. Gerstoft, "Semi-supervised source localization in reverberant environments with deep generative modeling," *IEEE Access*, vol. 9, pp. 84956–84970, 2021.

[62] H. Hammer, S. E. Chazan, J. Goldberger, and S. Gannot, "Dynamically localizing multiple speakers based on the time-frequency domain," *EURASIP J. Audio, Speech, Music Process.*, vol. 2021, no. 1, p. 16, Dec. 2021.

[63] P. Castellini, N. Giulietti, N. Falcionelli, A. F. Dragoni, and P. Chiariotti, "A neural network based microphone array approach to grid-less noise source localization," *Appl. Acoust.*, vol. 177, Jun. 2021, Art. no. 107947.

[64] P. Xu, E. J. G. Arcondoulis, and Y. Liu, "Acoustic source imaging using densely connected convolutional networks," *Mech. Syst. Signal Process.*, vol. 151, Apr. 2021, Art. no. 107370.

[65] N. Liu, H. Chen, K. Songgong, and Y. Li, "Deep learning assisted sound source localization using two orthogonal first-order differential microphone arrays," *J. Acoust. Soc. Amer.*, vol. 149, no. 2, pp. 1069–1084, Feb. 2021.

[66] S. Patel, M. Zawodniok, and J. Benesty, "DCASE 2020 task 3: A single stage fully convolutional neural network for sound source localization and detection," in *DCASE2020 Challenge*. IEEE, 2020.

[67] Z.-Q. Wang, X. Zhang, and D. Wang, "Robust speaker localization guided by deep learning-based time-frequency masking," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 27, no. 1, pp. 178–188, Jan. 2019.

[68] A. S. Subramanian, C. Weng, S. Watanabe, M. Yu, and D. Yu, "Deep learning based multi-source localization with source splitting and its effectiveness in multi-talker speech recognition," *Comput. Speech Lang.*, vol. 75, Sep. 2022, Art. no. 101360.

[69] W. He, P. Motlicek, and J.-M. Odobez, "Neural network adaptation and data augmentation for multi-speaker direction-of-arrival estimation," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 29, pp. 1303–1317, 2021.

[70] Y. Hao, A. Küçük, A. Ganguly, and I. M. S. Panahi, "Spectral flux-based convolutional neural network architecture for speech source localization and its real-time implementation," *IEEE Access*, vol. 8, pp. 197047–197058, 2020.

[71] T. N. T. Nguyen, K. N. Watcharasupat, N. K. Nguyen, D. L. Jones, and W.-S. Gan, "SALSA: Spatial cue-augmented log-spectrogram features for polyphonic sound event localization and detection," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 30, pp. 1749–1762, 2022.

[72] J. Ye, T. Kobayashi, and M. Murakawa, "Urban sound event classification based on local and global features aggregation," *Appl. Acoust.*, vol. 117, pp. 246–256, Feb. 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0003682X16302274

[73] J. Lee and J. Nam, "Multi-level and multi-scale feature aggregation using pretrained convolutional neural networks for music auto-tagging," *IEEE Signal Process. Lett.*, vol. 24, no. 8, pp. 1208–1212, Aug. 2017.

[74] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 818–833.

[75] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?" in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 1–9.

[76] C. Liu, L.-C. Chen, F. Schroff, H. Adam, W. Hua, A. L. Yuille, and L. Fei-Fei, "Auto-DeepLab: Hierarchical neural architecture search for semantic image segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 82–92.

[77] F. Isensee, P. F. Jäger, P. M. Full, P. Vollmuth, and K. H. Maier-Hein, "nnU-Net for brain tumor segmentation," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, vol. 12659. Cham, Switzerland: Springer, 2021.

[78] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent (MICCAI)*, 2015, pp. 234–241.

[79] Y. Han, Y. Wu, and Y. Jiang, "Feature aggregation network for image segmentation," in *Proc. 28th ACM Int. Conf. Multimedia*, 2020.

[80] M. Tan, R. Pang, and Q. V. Le, "EfficientDet: Scalable and efficient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10778–10787.

[81] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779–788.

[82] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Y. Fu, and A. C. Berg, "SSD: Single shot MultiBox detector," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 21–37.

[83] A. Odena, V. Dumoulin, and C. Olah, "Deconvolution and checkerboard artifacts," *Distill*, vol. 1, no. 10, p. 3, Oct. 2016.

[84] Y. Sugawara, S. Shiota, and H. Kiya, "Checkerboard artifacts free convolutional neural networks," *APSIPA Trans. Signal Inf. Process.*, vol. 8, no. 1, p. e9, 2019.

[85] Z. Wojna, V. Ferrari, S. Guadarrama, N. Silberman, L.-C. Chen, A. Fathi, and J. Uijlings, "The devil is in the decoder: Classification, regression and GANs," *Int. J. Comput. Vis.*, vol. 127, nos. 11–12, pp. 1694–1706, Dec. 2019.

[86] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.

[87] M. S. Brandstein and D. B. Ward, *Microphone arrays: Signal Processing Techniques and Applications*. Cham, Switzerland: Springer, 2001.

[88] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 936–944.

[89] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 8759–8768.

[90] G. Ghiasi, T.-Y. Lin, and Q. V. Le, "NAS-FPN: Learning scalable feature pyramid architecture for object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7036–7045.

[91] J. Redmon and A. Farhadi, "YOLOv3: An Incremental Improvement," in *Proc. Comput. Vis. Pattern Recognit.*, 2018.

[92] R. Caruana, "Multitask Learning," *Mach. Learn.*, vol. 28, no. 1, pp. 41–75, 1997.

[93] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2012, pp. 1097–1105.

[94] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2015.

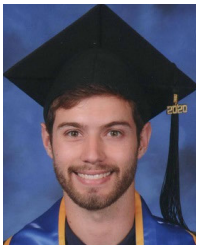
[95] S. Adavanne, J. Nikunen, A. Politis, and T. Virtanen, 2018, "TUT sound events 2018—Ambisonic, reverberant and real-life impulse response dataset," Tampere Univ. Technol., Finland, doi: 10.5281/zenodo.1237793.

[96] B. Healy, P. Mcnamee, and Z. N. Ahmadabadi. (Jun. 2023). *Feature Aggregation for Neural Network*. [Online]. Available: https://gitlab.com/dsim-lab/paper-codes/feature-aggregation-for-neural-networks

[97] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2015.

[98] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.

[99] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.
[100] J. Yin, Y. Zhang, Y. Liu, and W. Zhang, "DOA estimation for microphone array speech recognition based on LSTM recurrent neural network," in *Proc. Int. Conf. Acoust., Speech Signal Process.*, Sep. 2018.
[101] D. M. Powers, "Evaluation: From precision, recall and F-measure to ROC, informedness, markedness & correlation," *J. Mach. Learn. Technol.*, vol. 2, no. 1, pp. 37–63, 2011.
[102] A. Mesaros, T. Heittola, and T. Virtanen, "Metrics for polyphonic sound event detection," *Appl. Sci.*, vol. 6, no. 6, p. 162, May 2016.
[103] S. Adavanne, P. Pertilä, and T. Virtanen, "Sound event detection using spatial features and convolutional recurrent neural network," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 771–775.
[104] S. Adavanne, A. Politis, and T. Virtanen, "Direction of arrival estimation for multiple sound sources using convolutional recurrent neural network," in *Proc. 26th Eur. Signal Process. Conf. (EUSIPCO)*, Sep. 2018, pp. 1462–1466.
[105] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. Antennas Propag.*, vol. AP-34, no. 3, pp. 276–280, Mar. 1986.



BRENDAN HEALY received the Bachelor of Science degree in mechanical engineering from the University of California at Irvine, Irvine, CA, USA, in 2020, and the Master of Science degree in mechanical engineering from San Diego State University, San Diego, CA, USA, in 2023. His work experience includes research and development engineering for computer vision algorithms with Hologenix Inc., and robotic control software with RheoSense. At San Diego State University, he was with the Dynamic Systems and Intelligent Machines Laboratory, researching and deploying deep learning models for perception on mobile robots.

PATRICK MCNAMEE received the B.S. and M.Sc. degrees in aerospace engineering and the M.Sc. degree in computer science from the University of Kansas, Lawrence, KS, USA, in 2017, 2020, and 2021, respectively. From 2016 to 2021, he was a Undergraduate and Graduate Student Researcher with the University of Kansas, studying the dynamics and controls of unmanned aerial vehicles (UAVs) as well as machine learning applications for UAVs. He is currently a Graduate Student Researcher with the Dynamic Systems and Intelligent Machines Laboratory (DSIM), San Diego State University.

ZAHRA NILI AHMADABADI (Member, IEEE) was an Assistant Professor with Wichita State University. She is currently an Assistant Professor with the Mechanical Engineering Department and the Director of the Dynamic Systems and Intelligent Machines Laboratory, San Diego State University (SDSU). Her current research interests include acoustic perception, deep learning and sequence modeling, cooperative robotics, and nonlinear dynamical systems.