# CNN Models for Microphone Array Covariance Matrix Upsampling and Acoustic Imaging

Marianthi Adamopoulou*, Parthasaarathy Sudarsanam†, David Diaz-Guerra†, Meng Jiang*, Archontis Politis†, Seyed Jalaleddin Mousavirad*, Tuomas Virtanen†, Jan Lundgren*,
*Dept. of Comp. and Elec. Engineering Mid Sweden University Sundsvall, Sweden
†Audio Research Group, Tampere University Tampere, Finland
*Email: {name.lastname}@miun.se
†Email: {name.lastname}@tuni.fi

***Abstract*—Acoustic imaging visualization is a core methodology in acoustics, enabling spatial analysis of sound sources and acoustic scenes. However, limited sensor availability in practical systems motivate approaches that enhance spatial resolution without increasing the hardware complexity. In this paper, we focus on upsampling virtually a tetrahedral 4-microphone array to a spherical 32-microphone array by estimating the covariance matrices of the channels employing deep learning techniques. Five neural network architectures are investigated for covariance upsampling for acoustic imaging using the real-world STARSS23 dataset. These models are developed to estimate a 32-microphone, time–frequency covariance matrix from a 4-microphone input covariance representation. The proposed architectures are based on 2D convolutional layers to capture the underlying spatial–spectral structure of covariance matrices, and are further enhanced with frequency dynamic convolution to model their frequency-dependent properties. The proposed architectures are evaluated in terms of root mean square error (RMSE) and using delay-and-sum beamforming acoustic imaging. Quantitative results show that all models outperform a random-guess baseline, which yields an RMSE of 0.548, with the best-performing architecture achieving an RMSE of 0.432. We analyze qualitatively the performance of the proposed models through beamforming heatmap visualizations derived from the 4-channel input covariance, the 32-channel ground truth, and the predicted 32-channel covariance matrices. These results demonstrate that covariance upsampling significantly enhances the effective performance of the 4-channel microphone array, producing sound maps that closely resemble those obtained with the 32-channel array.**

***Index Terms*—Acoustic imaging, microphone array upsampling, spherical microphone arrays, sparse microphone arrays, spatial covariance matrix, beamforming, deep learning, convolutional neural networks**

## I. Introduction

Acoustic imaging visualizes the spatial distribution of sound sources in an environment by exploiting spatial information captured by microphone arrays followed by multichannel signal processing techniques. It is a key technology in instrumentation related to identification, characterization, and localization of sound sources in noisy and reverberant conditions. It is widely used in automotive and aerospace industries [1], [2], marine surveillance [3], monitoring of wind turbines [4], pipe leakages or electrical line discharges [5], and industrial machine monitoring or fault diagnosis [6]–[8].

Despite their effectiveness, acoustic imaging systems are fundamentally constrained by the size, geometry, and number of microphones in the array. Achieving high spatial resolution requires large-size dense arrays with many sensors, with increased hardware costs and computational complexity. In practical systems, such as mobile or embedded platforms, these requirements limit the scalability of the array. As a result, acoustic imaging with sparse or compact arrays often exhibits reduced spatial resolution and degraded localization performance, motivating methods that enhance spatial representations without increasing the physical array size and the number of microphones [9], [10]. Super-resolution approaches have been proposed that can produce acoustic images beyond the natural spatial resolution of the array. Those can be deconvolution-based [11], parametric [12], [13], spatial reassignment based [14], or based on sparse recovery [15]. Many of those methods rely on statistical dependencies between the microphones of the array, captured in the spatial covariance matrix (SCM) of the array signals.

In this work, we focus on a similar super-resolution task applied to spherical microphone arrays (SMAs), which offer a fully spherical field-of-view and uniform imaging resolution at all directions. Therefore, to address hardware limitations, a 4-channel microphone array along with deep-learning are employed to virtually upsample the array into a 32-channel SMA, by predicting the covariance matrices of the missing channels, aiming to improve the performance of the system. Acoustical imaging with SMAs is often performed in the spherical harmonic (SH) transform domain due to theoretical and computational advantages. Upsampling the SH transformed microphone signals to harmonic orders higher than the ones provided naturally by the SMA is, then, one way to spherical super-resolution imaging. In [16], [17], virtual microphone signals are generated by interpolating the real ones, increasing artificially the SH transform order of the array. In [15], [18] low-resolution SH signals are upscaled directly to higher-order ones by assuming spatial sparsity on the sound sources, while [19], [20] train deep-learning models for the same upscaling task.

An alternative approach to upscaling or interpolating the

The authors wish to acknowledge CSC—IT Center for Science, Finland, for computational resources.

array signals is to instead try to predict the interchannel signal statistics of a larger or denser virtual microphone array, in the form of its SCM, from the real observed SCM. Standard normal or super-resolution imaging methods can be then applied to such an upscaled SCM. In [21], the deep-learning image super-resolution model Deep Back Projection Network (DBPN) [22] is applied on the SCM of a 4-microphone tetrahedral array to estimate the SCM of a 32-microphone SMA, which is used to generate acoustic images using DeepWave, a deep-learning model that generates acoustical images from covariance matrices [23]. The replacement of DeepWave by a self-supervised model is later proposed in [24], but the upsampling of the SCM is still done with DBPN.

Even if both can be seen as 2D signals, covariance matrices are very different from the images DBPN was designed for. In images, the spatial distance between all the adjacent pixels is supposed to be the same, which is not the case for inter-channel covariance matrices, where the distance between elements depend on the array geometry and the order in which the microphones have been numbered. This means that using 2D convolutions over the covariance matrices, as DBPN does when applied to them, might not be and optimal choice. In addition, if DBPN is applied independently to the covariance matrices of each time-frequency bin as done in [21], [24], the model cannot exploit the correlations between different bins.

In the present work, we propose a deep learning framework for SCM upsampling that operates directly on time–frequency dependent SCMs, preserving both temporal and frequency-specific spatial structure. As in [21], [24], our method maps low-dimensional SCM representations from sparse microphone arrays to high-dimensional SCMs, enabling general-purpose acoustic imaging. However, we use models specifically designed for covariance matrices: 2D convolutional layers are applied across the time and frequency dimensions and the real and imaginary parts of the non-redundant elements of the matrices are treated as convolution channels. We train and evaluate the proposed approach on the real-world STARSS23 dataset and conduct a systematic comparison of convolutional neural network architectures, including frequency-dependent convolutional layers, to analyze their impact on SCM reconstruction performance.

## II. Methodology

In this work, we study and compare five different convolutional neural networks (CNN) to upsample the SCM of a microphone array with a reduced number of microphones into the one that would have been obtained with an array with more microphones. Specifically, we tackle the case of upsampling the SCM of a tetrahedral array of 4 microphones to the one obtained by a 32-microphone spherical array, though the same methodology could be used with different arrays. We use the convolutions to incorporate time-frequency context to the computation of every covariance matrix, with the matrix itself not being treated as an image but as convolution channels. Therefore, we do not impose any prior assumptions about the array geometry into the model.

### A. Input representation

Let $X_m(t, f)$ be the short-time Fourier transform (STFT) of the signal captured by the $m$-th microphone of the array, which we compute using a Hann window with an FFT length of 512 samples, a 50% hop size, and a sampling rate of 24 kHz. For both the 4-microhpone and the 32-microphone arrays, we can construct the instantaneous SCM of each time-frequency bin as

$$\mathbf{C}_x(t, f) = \mathbf{X}(t, f)\mathbf{X}^H(t, f) \tag{1}$$

where $\mathbf{X}(t, f) = [X_1(t, f), \ldots, X_M(t, f)]^\top$ is the vector of STFT coefficients across the $M$ microphones, and $(\cdot)^H$ denotes the Hermitian transpose. These instantaneous SCMs capture the inter-channel spatial structure of the sound field at each time-frequency bin. To reduce noise and smooth the covariance matrix, the instantaneous SCMs are averaged over short temporal windows, producing 5 SCMs per second. Finally, we normalize the SCM of each time-frequency bin by dividing it by the value of the first element of its diagonal (i.e., the energy in the first sensor at that time-frequency bin) for an easier processing by the models.

Since the SCMs are Hermitian matrices with real-valued diagonal, they only contain only $M^2$ non-reductant values: the real part of the diagonal elements and the real and the imaginary parts of the elements bellow the diagonal. These components are stacked as convolution channels to form a $\mathbb{R}^{C\times F\times T}$ tensor, where $C = M^2$ is the number of non-reductant covariance channels, $F$ is the number of frequency bins, and $T$ is the number of time frames. This tensor, computed from the signals of the tetrahedral and the spherical array, serves as the input and groundtruth output of the model respectively.

### B. Neural Network Architectures

We investigate the CNN architectures for upsampling spatial covariance representations, as illustrated in Fig. 1. All models are designed to map a low-dimensional covariance representation obtained from a sparse 4-channel array to a high-dimensional representation corresponding to a 32-channel array. They are trained in a supervised learning setting by minimizing the mean squared error (MSE) between the predicted and target SCMs. [1]

*a) **Base CNN**:* The base CNN is composed of a sequence of six 2D convolutional blocks with kernel size $3 \times 3$ and progressively increasing channel dimensionality as shown in Figure 1 (a). Starting from 16 input channels, corresponding to the flattened 4-channel covariance representation, the network gradually increases the number of channels until reaching 1024 output channels, corresponding to the 32-channel covariance matrix. Each convolutional block consists of a 2D conv layer, which is employed for spectro-temporal analysis of the data, followed by channel normalization, to accelerate the training, and a ReLU activation function, while dropout is applied in the deeper layers to reduce overfitting.

[1] All employed codes are publicly available in: https://github.com/marianthiadm/Upsampling-sparse-microphone-array-with-CNN

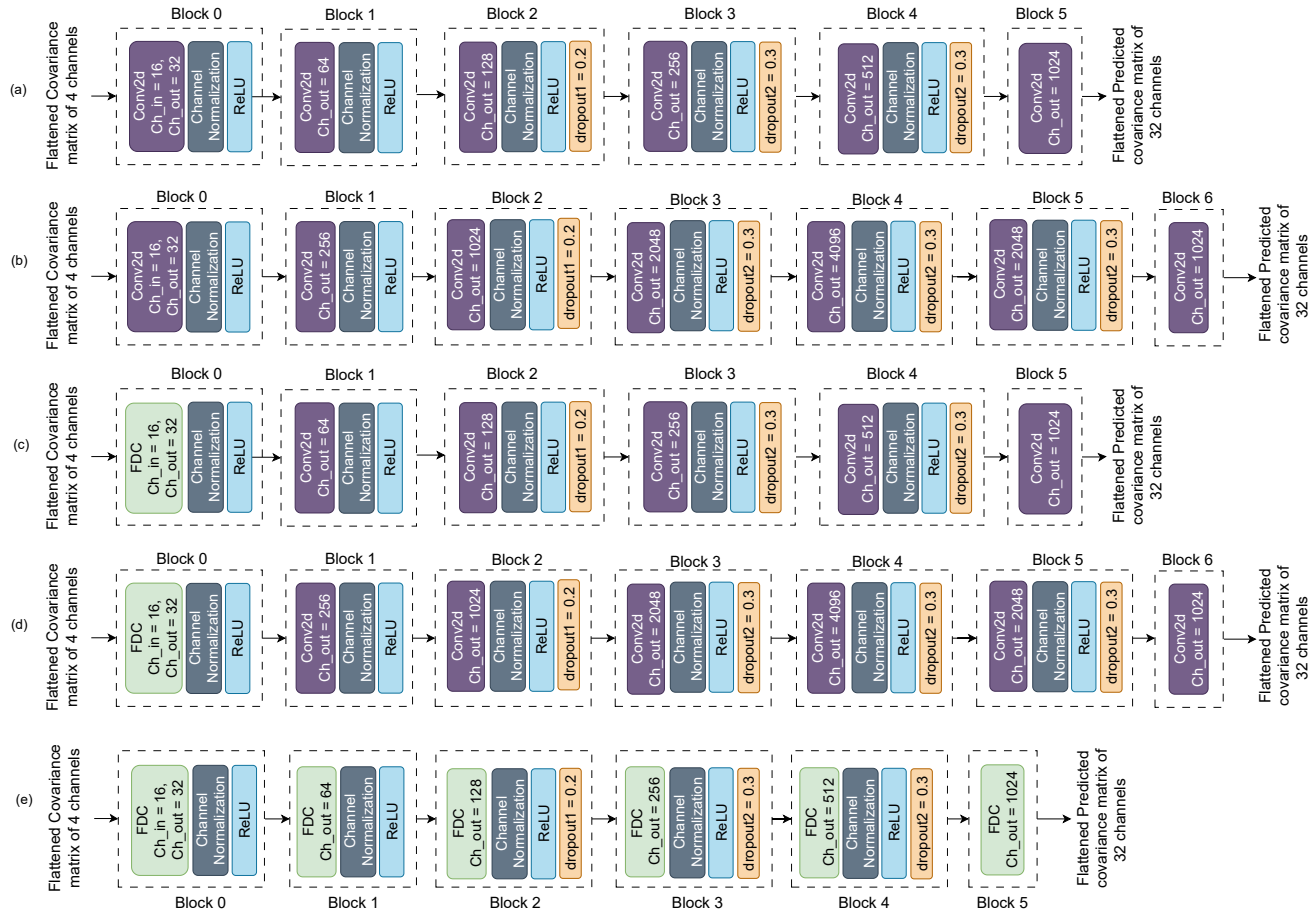


Fig. 1: model architectures. (a) Base CNN, (b) Expanded CNN, (c) Hybrid FDC-CNN Base, (d) Hybrid FDC-CNN Expanded, (e) Full FDC-CNN

*b)* ***Expanded CNN****:* The expanded CNN architecture extends the base CNN by increasing the representational capacity of the network as shown in Figure 1 (b). It follows the same overall structure and layer sequence as the base CNN, but it introduces an additional channel expansion stages in which the channel dimension is increased beyond the target size. Specifically, the network expands the channel dimension up to 4096 channels before projecting it back to the required 1024 output channels. This overparameterization is intended to allow the network to learn richer intermediate spatial representations before reconstruction.

*c)* ***Hybrid FDC-CNN Base****:* The previous models use the same convolutional kernels to process all the frequency bins, which might not be optimal considering that the SCMs are frequency dependent. To allow the models to process each frequency differently, we study the use of frequency dependent convolutions (FDC) implemented as 1D convolutions across time with a different kernels for each frequency bin. In a similar way to [25], the Hybrid FDC-CNN base model is equal to the base CNN just replacing the first convolutional layer by a FDC as shown in Figure 1 (c).

*d)* ***Hybrid FDC-CNN Expanded****:* This architecture also uses an FDC layer as shown in Figure 1 (d). Similar to the Expanded CNN, it follows the exact same architecture, where the only difference is found in the first block. Thus, here as earlier, the standard CNN layer of the first block is replaced with an FDC layer. Contentiously, the remaining blocks use standard convolutional layers with channel normalization, ReLU activation, and dropout as the expanded model.

*e)* ***Full FDC-CNN****:* The full FDC-CNN architecture further extends the frequency-aware design by employing FDC layers in all the convolutional blocks as shown in Figure 1 (e). This model is the most computationally demanding among the proposed architectures, as every layer dynamically adapts its kernels to frequency-specific characteristics. Consequently, the extension was implemented on the base architecture, as the expanded model encompasses high complexity and training time. This fully frequency-adaptive design aims to maximize the model's ability to capture fine-grained spectral–spatial dependencies in the covariance domain.

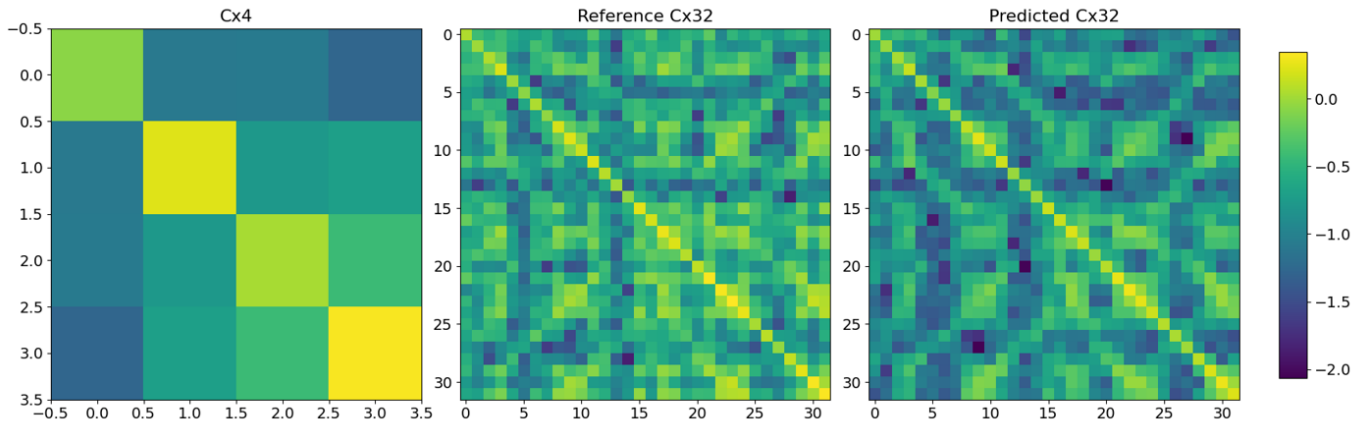


Fig. 2: Frequency average of the SCMs for a single time frame. Left is the 4-channel input SCM, in the middle is the ground truth of the 32-channel SCM, and right is the estimation of the Hybrid FDC-CNN Expanded model.

### C. Dataset

All models were trained on the development–train split of the Sony-TAu Realistic Spatial Soundscapes 2023 (STARSS23) dataset [26], provided as part of DCASE 2024 Task 3. The dataset consists of multichannel recordings of real acoustic scenes captured in a variety of rooms and environments using a 32-channel EigenMike microphone array. The STARSS23 development dataset contains approximately 7.5 hours of recordings collected across 16 different rooms and includes 13 sound event classes, such as female and male speech, clapping, telephone, laughter, domestic sounds, footsteps, doors, music, musical instruments, water taps, bells, and knocking sounds [27]. For the purposes of this study, the dataset has been segmented into 5-second clips to form the input of the models. It should be noted that certain segmented audio clips were affected by saturation, while others exhibited substantially higher energy over one channel; to avoid bias these segments were removed.

Model evaluation was performed on the STARSS23 public test dataset, which was not used for early stopping nor did it influence the training result in any way. The 4 channels that are used for the model input have been selected from the 32-channel EigenMike spherical array to approximate a tetrahedral array and correspond to those available in the public version of the dataset [27].

### D. Training

All models are trained for 100 epochs using the Adam optimizer with a learning rate of $1 \times 10^{-4}$. We did not conduct any extensive hyperparameter optimization, so more optimal configurations might exist. Table I shows the GPU memory usage during training, and it is observed that replacing the first layer of the models with an FDC layer does not affect memory usage. On the contrary, when building the entire model with FDC layers, GPU memory usage doubles, as does the number of parameters. It is worth noting that the training process varies between around $3\,\text{hours}$ and $10\,\text{days}$ for the base CNN architecture and the full FDC-CNN architecture, respectively. The evaluation, which includes the computation of the 32-channel SCMs for more than $3\,\text{hours}$ of audio, the model inference, and the RMSE computation, takes approximately $1\,\text{hour}$ and $20\,\text{minutes}$ for all architectures.

## III. Results

To quantitatively assess model performance, we compute the root mean squared error (RMSE) between the predicted 32-channel SCMs and the corresponding ground-truth SCMs on the STARSS23 dev-test dataset and average them across frequency. As a reference, we also include the RMSE of the random guess solution, which we define as the average $\mathbf{C}_x(t, f)$ of the dev-train groundtruths. As we can see in Table I, there are not big differences between models, but they all clearly outperform the random guess, proving that they are not only exploiting dataset biases, but they are able to extract information from the 4-channel input SCM.

The use of architectures with larger number of convolutional channels seems to have a bigger impact than the use of FDCs, though using FDCs at the first layer of the models provides some minor improvements of the performance with a negligible impact in the model size. Contrary to this, replacing all the convolutional layers by FDCs radically increases the model size and has a negative impact on the results, probably

TABLE I: Number of parameters, peak GPU memory consumption during training and RMSE value for each model

| Model architecture | Parameters | GPU memory | RMSE |
|---|---|---|---|
| Random guess | | | 0.548 |
| Base CNN | 6.29 M | 2.03 GB | 0.452 |
| Expanded CNN | 191.29 M | 4.81 GB | 0.433 |
| Hybrid FDC-CNN Base | 6.69 M | 2.04 GB | 0.447 |
| Hybrid FDC-CNN Expanded | 192.09 M | 4.82 GB | 0.432 |
| Full FDC-CNN | 539.36 M | 10.13 GB | 0.451 |

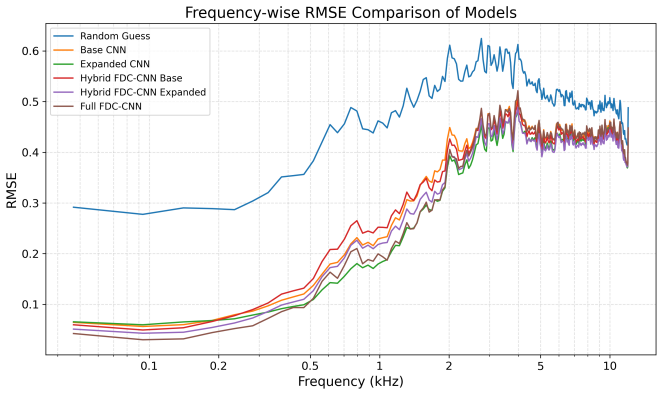


Fig. 3: RMSE loss per frequency (log. scale) for each proposed model and the random guess.

because of not having enough data to exploit higher power of a bigger model.

To analyze the frequency-dependent behavior of the upsampling process, the RMSE of the SCMs of each frequency bin are additionally reported in Fig. 3. As could be expected, the estimation error grows with frequency till reaching the aliasing limit of the 4-channel array used as input for the model. However, even after exceeding that limit, the models are still able to provide better estimates than the random guess. This is probably because the models are able to use the information of the lower frequencies to predict the high frequency SCMs.

In Fig. 2, three SCMs of a dataset sample, are presented. To the left it is shown the 4-channel input SCM, in the middle is shown the ground truth 32-channel SCM, and to the right is shown the 32-channel SCM predicted by the Hybrid FDC-CNN Expanded architecture; all of them are averaged across all frequency bins, for a specific time frame. We can see how the patterns present in the groundtruth seem to be intensified in the estimation of the model, which looks like a denoised and sharpened version.

For qualitative evaluation, delay-and-sum beamforming is applied to the input, ground truth, and predicted SCMs. This computation is based on the analytical solution for the transfer function of a rigid sphere [2]. Finally, the produced heatmaps are overlaid on the corresponding video recordings to visually assess spatial reconstruction accuracy. Despite we focus on delay-and-sum beamforming, more advanced techniques could be employed to generate acousit images from the upsampled SCMs, including those based on deep learning such as DeepWave [23] and LAM [24].

Fig. 4 (a) presents the heatmap corresponding to the 4 channels input data, while Fig. 4 (b) shows the heatmap of the predicted 32 channels and Fig. 4 (c) depicts the heatmap for the ground truth 32 channels. In this example, the sound event corresponds to a ringing mobile phone and its spatial reflections. Therefore, it is worth to mention that the 32-

[2]https://github.com/polarch/Array-Response-Simulator/blob/master/simulateSphArray.m

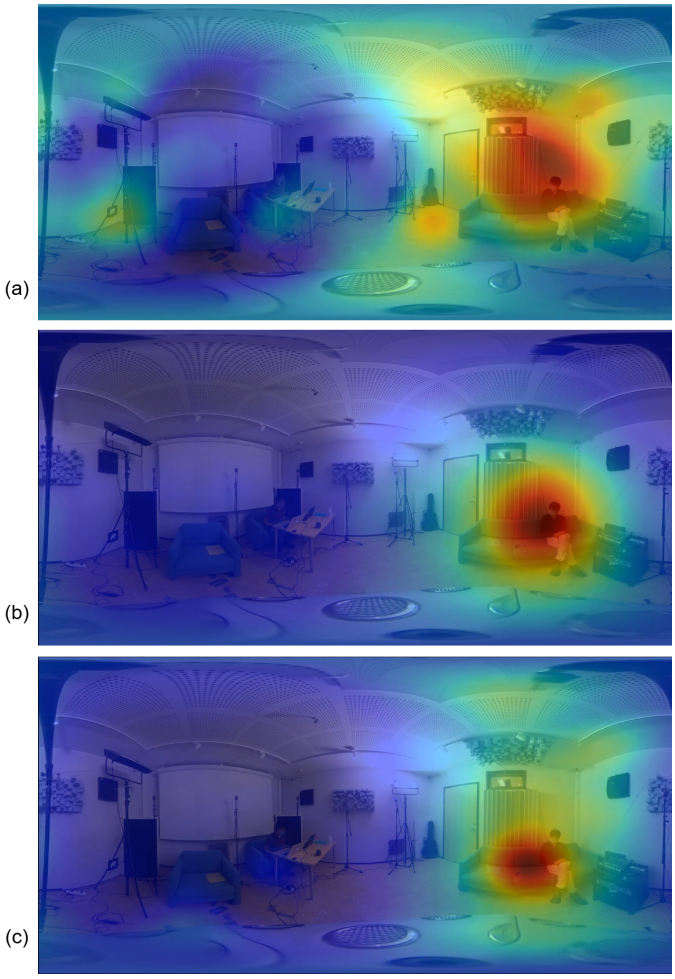


Fig. 4: Beamformed sound heatmaps for (a) 4-channel, (b) predicted 32-channel and (c) reference 32-channel

channel original audio accurately localizes the sound source, whereas the predicted map is slightly shifted upwards, but it is still more accurate than the map from the original 4-channel input data. We can also observe how the heatmap generated from the estimated 32-channel SCM is more focused on the main audio source and ignores the ceiling and wall reflections present in the groundtruth. This is probably a consequence of the sharpening observed on the covariance matrices.

Despite being training to replicate the SCM of a 32-channel array without any kind of further processing, the models seem to focus on the stronger sources and filter out the wall reflections. This is probably due to the stronger sources being easier to find and model by the models, and could be an advantage or a disadvantage depending on the final application of the system.

## IV. Conclusion

This study compared five CNN model architectures for virtual microphone array upsampling from a tetrahedral 4-channel array to a spherical 32-channel, emphasizing on sound event localization and spatial distribution. Performance evalu-

ation using RMSE and conventional beamforming highlights the challenge of direct upsampling from 4-channel covariance matrix to 32-channel, given the sparse spatial information. The Hybrid FDC-CNN Expanded architecture accomplished the lowest RMSE value of 0.432. All proposed models achieve better performance for low frequency over high frequency bins, due to aliasing limit of the 4-channels. However, each model consistently outperforms the random guess baseline at all frequencies, affirming its ability to learn spatial characteristics, successfully localizing sound events. In comparison to the ground truth, the predicted heatmaps are more concentrated to the sound sources while ignoring to model reflections and reverberations. The models with FDC layers learn frequency dependent characteristics and with more modeling capacity, it performs better than the standard CNN models. Nevertheless, the optimal model choice depends on application-specific requirements, taking into considerations the model complexity and frequency dependent characteristics, balancing performance and memory size of the implemented system. Future works will focus on reducing the computational complexity while maintaining or improving the spatial distribution on the produced acoustic maps.

## REFERENCES


[1] T. Rittenschober and R. Karrer, "High-speed acoustic imaging for the localization of impulse-like sound emissions from automotive components," in *13th International Styrian Noise, Vibration & Harshness Congress: The European Automotive Noise Conference*, 2024.

[2] R. Merino-Martínez, P. Sijtsma, M. Snellen, T. Ahlefeldt, J. Antoni, C. J. Bahr, D. Blacodon, D. Ernst, A. Finez, S. Funke *et al.*, "A review of acoustic imaging methods using phased microphone arrays: Part of the "aircraft noise generation and assessment" special issue," *CEAS Aeronautical Journal*, vol. 10, no. 1, pp. 197–230, 2019.

[3] M. S. Dantzker, M. T. Duggan, E. Berlik, S. Delikaris-Manias, V. Bountourakis, V. Pulkki, and A. N. Rice, "Deciphering complex coral reef soundscapes with spatial audio and 360° video," *Methods in Ecology and Evolution*, vol. 16, no. 11, pp. 2622–2637, 2025.

[4] S. Ding, C. Yang, and S. Zhang, "Acoustic-signal-based damage detection of wind turbine blades—a review," *Sensors*, vol. 23, no. 11, p. 4987, 2023.

[5] J. Pihera, J. Hornak, P. Trnka, O. Turecek, L. Zuzjak, K. Saksela, J. Nyberg, and R. Albrecht, "Partial discharge detection using acoustic camera," in *2020 IEEE 3rd International Conference on Dielectrics (ICD)*. IEEE, 2020, pp. 830–833.

[6] M. Orman and C. T. Pinto, "Usage of acoustic camera for condition monitoring of electric motors," in *2013 IEEE International Conference of IEEE Region 10 (TENCON 2013)*. IEEE, 2013, pp. 1–4.

[7] Z. Xu and Z. Chen, "Investigation of noise characteristics of electric vibrator utilizing acoustic camera and transfer path analysis," *Measurement Science and Technology*, vol. 35, no. 11, p. 116012, 2024.

[8] W. Fiebig and D. Dabrowski, "Use of acoustic camera for noise sources localization and noise reduction in the industrial plant," *Archives of Acoustics*, vol. 45, no. 1, pp. 111–117, 2020.

[9] T. Lübeck, J. M. Arend, and C. Pörschmann, "Spatial upsampling of sparse spherical microphone array signals," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 1163–1174, 2023.

[10] F. Miotello, F. Terminiello, M. Pezzoli, A. Bernardini, F. Antonacci, and A. Sarti, "A physics-informed neural network-based approach for the spatial upsampling of spherical microphone arrays," in *2024 18th International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2024, pp. 215–219.

[11] T. F. Brooks and W. M. Humphreys, "A deconvolution approach for the mapping of acoustic sources (damas) determined from phased microphone arrays," *Journal of sound and vibration*, vol. 294, no. 4-5, pp. 856–879, 2006.

[12] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE transactions on antennas and propagation*, vol. 34, no. 3, pp. 276–280, 1986.

[13] R. P. Dougherty, "Functional beamforming," in *5th Berlin beamforming conference*. GFaI, eV Berlin, 2014, pp. 19–20.

[14] L. McCormack, A. Politis, and V. Pulkki, "Sharpening of angular spectra based on a directional re-assignment approach for ambisonic sound-field visualisation," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 576–580.

[15] N. Epain, C. Jin, and A. Van Schaik, "The application of compressive sampling to the analysis and synthesis of spatial sound fields," in *Audio Engineering Society Convention 127*. Audio Engineering Society, 2009.

[16] C. D. Salvador, S. Sakamoto, J. Treviño, and Y. Suzuki, "Enhancing binaural reconstruction from rigid circular microphone array recordings by using virtual microphones," in *Audio Engineering Society Conference: 2018 AES International Conference on Audio for Virtual and Augmented Reality*. Audio Engineering Society, 2018.

[17] T. Lübeck, J. M. Arend, and C. Pörschmann, "Spatial upsampling of sparse spherical microphone array signals," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 1163–1174, 2023.

[18] A. Wabnitz, N. Epain, A. McEwan, and C. Jin, "Upscaling ambisonic sound scenes using compressed sensing techniques," in *2011 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2011.

[19] G. Routray, S. Basu, P. Baldev, and R. M. Hegde, "Deep-sound field analysis for upscaling ambisonic signals," in *EAA Spatial Audio Signal Processing Symposium*, 2019.

[20] E. Chatzimoustafa and P. Jax, "Higher-order ambisonics upscaling using gated recurrent units," in *2025 33st European Signal Processing Conference (EUSIPCO)*, 2025.

[21] A. S. Roman, I. R. Roman, and J. P. Bello, "Robust doa estimation from deep acoustic imaging," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 1321–1325.

[22] M. Haris, G. Shakhnarovich, and N. Ukita, "Deep back-projection networks for super-resolution," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[23] M. Simeoni, S. Kashani, P. Hurley, and M. Vetterli, "Deepwave: a recurrent neural-network for real-time acoustic imaging," *Advances In Neural Information Processing Systems*, vol. 32, 2019.

[24] A. S. Roman, I. R. Roman, and J. P. Bello, "Latent acoustic mapping for direction of arrival estimation: A self-supervised approach," in *2025 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2025.

[25] M. Heikkinen, A. Politis, and T. Virtanen, "Neural ambisonics encoding for compact irregular microphone arrays," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 701–705.

[26] K. Shimada, A. Politis, P. Sudarsanam, D. A. Krause, K. Uchida, S. Adavanne, A. Hakala, Y. Koyama, N. Takahashi, S. Takahashi *et al.*, "STARSS23: An audio-visual dataset of spatial recordings of real scenes with spatiotemporal annotations of sound events," *Advances in Neural Information Processing Systems*, vol. 36, pp. 72 931–72 957, 2023.

[27] Diaz-Guerra, A. Politis, K. S. Parthasaarathy Sudarsanam, D. A. Krause, K. Uchida, N. T. Yuichiro Koyama, S. Takahashi, T. Shibuya, Y. Mitsufuji, and T. Virtanen, "Baseline models and evaluation of sound event localization and detection with distance estimation in DCASE2024 challenge," in *Workshop on Detection and Classification of Acoustic Scenes and Events*, 2024.