Head-Related Transfer Function Upsampling Using an Autoencoder-Based Generative Adversarial Network with Evaluation Framework

Xuyi Hu,[1] **Jian Li,**[1] **Lorenzo Picinali,**[1] *AES Member,* **AND Aidan O. T. Hogg**[1, 2]
(xh519@ic.ac.uk) (jl2622@ic.ac.uk) (l.picinali@imperial.ac.uk) (a.hogg@qmul.ac.uk)

[1] *Audio Experience Design - www.axdesign.co.uk, Dyson School of Design Engineering, Imperial College London, UK*
[2] *Centre for Digital Music, School of Electronic Engineering and Computer Science, Queen Mary University of London, UK*

Accurate Head-Related Transfer Functions (HRTFs) are essential for delivering realistic 3D audio experiences. However, obtaining personalised, high-resolution HRTFs for individual users is a time-consuming and costly process, typically requiring extensive acoustic measurements. To address this, spatial upsampling techniques have been developed to estimate high-resolution HRTFs from sparse, low-resolution acoustic measurements. This paper presents a novel approach leveraging the spherical harmonic (SH) domain and an Autoencoder Generative Adversarial Network (AE-GAN) to tackle the HRTF upsampling problem. Comprehensive evaluations are conducted using both perceptual models and objective spectral metrics to validate the accuracy and realism of the upsampled HRTFs. The results show that the proposed approach outperforms traditional barycentric interpolation in terms of log-spectral distortion (LSD), particularly in extreme sparsity scenarios involving fewer than 12 measurements. These results go some way to justifying that the proposed AE-GAN approach is able to create high-quality, high-resolution HRTFs from only a few acoustic measurements, helping pave the way for more accessible personalised spatial audio across a range of applications.

0 INTRODUCTION

Advancements in spatial audio technology have enabled significant improvements in virtual and augmented reality (VR/AR), immersive gaming, and hearing assistive devices [1]. A key component of these advancements is the ability to accurately reproduce immersive audio, which allows users to experience sound as if it were coming from different directions [2], mimicking what we hear in the real world [3]. Central to this effort is the use of Head-Related Transfer Functions (HRTFs), which capture the unique filtering effects caused by an individual's anatomy, including the head, torso, and pinnae, on incoming sound waves [4]. Adapting these HRTFs to individual listeners is still a significant active research area. This has led to extensive studies on HRTFs, which characterize the listener-specific filtering effects introduced by anatomical structures. These effects arise as sound waves reflect and scatter off the head, torso, and pinnae before reaching the ear canal. HRTFs encode both interaural differences (i.e. disparities in the signals received by each ear) and monaural localization cues [5].

Research has shown that using non-individualized HRTFs can significantly impair sound source localization accuracy [6–8], as spectral cues are highly dependent on a listener's unique anatomy, particularly the shape of their pinnae [9]. Beyond localization, non-individualized HRTFs can also degrade perceptual attributes such as externalization, immersion, coloration, realism, and depth perception [10–12]. Moreover, the choice of HRTF can greatly influence a listener's ability to understand speech in complex auditory environments, such as cocktail party scenarios [13].

Various approaches have been explored for HRTF individualization, including direct acoustic measurements [14], 3D scanning [15], morphological modeling of ear geometry [16, 17], and selecting the best-fitting HRTF from a database of prerecorded measurements. Selection methods typically rely on either morphology-based techniques [18, 19] or perceptual-based evaluations, such as listener preference [20] or localization accuracy [21, 22]. A comprehensive overview of these techniques can be found in [23].

Among these methods, direct acoustic measurements [24] remain the gold standard. However, this approach requires a specialized and expensive setup and is time-consuming, as it involves capturing hundreds to thousands of impulse responses (IRs) across different spatial positions [14]. Techniques such as interlaced sine sweeps [25] can accelerate the process, particularly for elevation measurements, but do not fully address the issues regarding the time required. Other methods aimed at improving measurement efficiency [26, 27] often demand high-end equipment, making them impractical for widespread adoption.

As a result, researchers have explored alternative methods such as 3D scans [15], geometric modelling [16, 17] and database matching [18, 19, 21, 22]. An overview of some of the most common methods can be found in [23]. While these methods have achieved varying degrees of success, they often underperform compared to acoustic measurements.

Another promising approach to address the difficulties attached to acoustically measuring an HRTF is spatial upsampling, where high-resolution HRTFs (typically containing over 300 IRs from multiple directions) are estimated from sparse, low-resolution measurements (which include only a few IRs from limited directions) [28]. This reduces the time required, the number of speakers needed and the possible need for the subject or speakers to be rotated during the HRTF measurement [29].

There are two primary categories of HRTF upsampling methods: algorithmic and learning-based. Algorithmic approaches rely on interpolation, constructing HRTFs for new source positions by superimposing existing HRTFs or basis functions derived from them. This interpolation is typically performed independently for each frequency or time sample using only a sparse HRTF dataset or even a subset of it. Learning-based approaches, in contrast, generate HRTFs for new source positions using neural networks trained on high-resolution datasets. These models learn relationships between HRTFs at different source positions, frequencies, and time samples, potentially leading to more accurate upsampling. However, algorithmic approaches work with almost any set of source positions, whereas neural networks are typically trained for specific spatial configurations. Additionally, neural networks can introduce hallucination errors, whereas well-parameterized algorithmic methods are more robust to unseen data.

One of the most widely used algorithmic methods for HRTF upsampling is barycentric interpolation [30–32]. This method performs well when the HRTFs dataset is relatively dense (e.g. with measurements spaced 10–15° apart) [33]. However, its reliability decreases when interpolating sparse measurements (e.g. spaced 30–40° apart). Another common approach is spherical harmonic (SH) interpolation [34–38], which similarly struggles with sparse input data. This is because these methods rely on averaging existing data points based on prior assumptions. For instance, barycentric interpolation uses the three nearest neighbours to compute a weighted average, but as the spacing between neighbours increases, upsampling accuracy declines.

Recently, machine learning (ML) methods have gained traction in HRTF personalisation research. Previous studies have demonstrated that ML techniques can estimate HRTFs from a listener's anthropometric measurements. For example, [39] employed a DNN-based model to synthesize personalized HRTFs using user-specific anthropometric features, achieving a log-spectral distortion (LSD) of 3.2 dB. This method utilized an autoencoder to reduce the dimensionality of raw HRTFs, preventing overfitting due to the typically small dataset size. The autoencoder's decoder then estimated HRTF magnitudes using the latent representation produced by a DNN trained on anthropometric features and target azimuth. Building on this approach, [40] introduced a dual-autoencoder model: one for compressing azimuth and anthropometric features and another for reducing the dimensionality of full HRTF magnitudes. This approach achieved an LSD of 4.3 dB.

ML has also been applied to HRTF upsampling. [41] proposed a method extending a regularised linear regression (RLR) approach based on the spherical wavefunction decomposition [42]. This method separated HRTFs into source position-dependent and source position-independent components, utilizing an autoencoder conditioned on source positions. It achieved an LSD of 4.4 dB when upsampling from 9 to 440 positions. Other ML-based methods include [43], which used a deep belief network (DBN) to achieve an average LSD below 3 dB, though only for upsampling from 125 to 1250 positions—still a relatively dense scenario. Another approach in [44] employed a convolutional neural network (CNN) demonstrating strong performance with LSD values of 4.4 dB for upsampling from 23 to 1250 positions and 3.8 dB for 105 to 1250 positions. However, this model processed HRTF data as a set of 2D slices rather than considering the full spherical representation. Additionally, ML techniques have been used in combination with the spherical harmonic transform (SHT) interpolation as a postprocessing step [45].

In 2024, in recognition of the need for standardized benchmarking, the SONICOM IEEE SPS Listener Acoustic Personalisation (LAP) Challenge was undertaken, which introduced a common evaluation framework for spatial upsampling techniques [46]. The challenge provided a rigorous testbed for evaluating HRTF upsampling methods across different sparsity levels. The results from the challenge highlighted the strengths and limitations of existing methods, particularly emphasizing the need for evaluations using the perceptual auditory models. In light of the LAP challenge, in this paper, we not only evaluate the proposed method using the LAP challenge's standardized benchmarking on objective metrics but also include a perceptual model evaluation. We also include a comparison with the challenge baselines as well as two of the eight challenge submissions.

In this paper, we build on our novel approach presented in [47]. This approach aims to leverage ML and SHs by using an Autoencoder Generative Adversarial Network (AE-GAN) deployed in the SH domain to tackle the problem of HRTF upsampling. The primary contributions of this study are as follows:

1. A modified AE-GAN architecture, based on the architecture presented in [47], that uses Bayesian optimization for hyperparameter tuning, leading to a more efficient and stable training process.
2. A new study into optimal speaker placement for acoustically measuring a low-resolution HRTF for upsampling.
3. An experimental study including comparisons with state-of-the-art methods such as SUpDEQ-MCA [35, 38] and GEP-GAN [29, 48] both of which were presented in the LAP Challenge [46].
4. Perceptual auditory model evaluations of the proposed AE-GAN method against traditional interpolation and other state-of-the-art techniques.

This paper is structured as follows: Section I introduces the method, including the pre-processing steps along with the model architecture. Section II explains the experimental setup, including the dataset used and how the model is trained. Section III explores model optimisation, and a study of optimal speaker placement is performed. In Section IV, spectral and perceptual model-based evaluations are presented. Finally, Section V provides the conclusions drawn.

1 METHOLOGY

1.1 Data Pre-Processing

An individual's HRTF data is typically represented by a set of data points distributed non-uniformly across the surface of a sphere. CNNs can be employed to extract features from data with spatial information. However, the limitation of CNNs is that they are most effective when dealing with data with uniform spacing. Applying CNNs directly to the raw HRTF data introduces challenges with aligning the convolutional kernels to the data, as the fixed convolutional kernel size and stride assume that each data point has neighbouring points at regular, predictable intervals. Since HRTF data are not uniformly distributed, this assumption breaks down, leading to misalignment and ineffective feature extraction. Thuillier et al. [49] proposed spherical CNNs that leverage neural processes to learn and predict HRTFs at arbitrary points on a sphere, addressing the challenges of sparse and irregularly sampled HRTF data. However, Implementing neural process meta-learners can be computationally demanding.

In this work, the SH transformation is adopted in data preprocessing for its significant advantages. This approach not only circumvents the challenge of adapting CNNs to the non-uniform nature of HRTF data but also improves computational efficiency. The SH coefficient F_l^m of degree l and order m is defined as:

$$F_l^m = \int_0^{2\pi}\int_0^{\pi} f(\theta,\phi)Y_l^m(\theta,\phi)\sin(\phi)d\phi d\theta\,, \tag{1}$$

where $Y_l^m(\theta,\phi)$ is the SH basis function of degree l and order m. θ and ϕ represent the azimuth and elevation angles, respectively. $f(\theta,\phi)$ is the original HRTF data function defined on the sphere. In acoustics, the SH basis function is defined as:

$$Y_l^m(\theta,\phi) = \sqrt{\frac{(2l+1)(l-m)!}{4\pi(l+m)!}}P_l^m(\cos(\phi))e^{jm\theta}\,, \tag{2}$$

where $P_l^m(x)$ are the associated Legendre functions. The inverse SHT is a process that reconstructs the original HRTF function from its SH coefficients F_l^m. The formula for inverse SHT is given by:

$$f(\theta,\phi) = \sum_{l=0}^{\infty}\sum_{m=-l}^{l} F_l^m Y_l^m(\theta,\phi)\,. \tag{3}$$

By undergoing the SHT, the original spatial HRTF data are decomposed into a series of coefficients, each of which represents a unique sound energy distribution pattern in the space. This decomposition captures essential spatial features and significantly reduces the complexity of the raw HRTF data. As a result, the data can be represented in a smooth, continuous form, enabling efficient interpolation and upsampling at arbitrary points on the sphere. The proposed procedure of upsampling is as follows: sparse measurements of HRTFs are first transformed into low-resolution SH coefficients. Next, the GAN-based model upsamples the low-resolution coefficients. The upsampled high-resolution SH coefficients produced by the GAN are then converted back into high-resolution HRTFs using the inverse SHT.

1.2 Model Structure

Our proposed method builds upon and extends the AE-GAN architecture presented in our previous DAFx paper [47], introducing improvements in discriminator design, training strategy, and spatial conditioning to enhance performance across sparse HRTF grids.

1.2.1 Autoencoder

The generator network is an autoencoder with an encoder-decoder structure. The primary goal of the encoder is to analyze the low-resolution SHs and find the latent representation z which contains the most salient features. The decoder aims to project this latent representation to a higher dimension, generating high-resolution coefficients.

As shown in Fig. 1a, the encoder is primarily constructed using a sequence of concatenated residual blocks. Such a structure is used to effectively extract features from low-resolution coefficients while avoiding the vanishing gradient problem. Two fully connected layers at the output stage are employed to compress the feature map into a lower-dimensional latent space, obtaining the latent representation z. Batch normalization is incorporated throughout the encoder to stabilize the training process and serves as a regularization technique. The parametric rectified linear unit (PReLU) is utilized as the activation function to introduce non-linearity. The PReLU function is mathematically expressed as:

$$\mathbf{PReLU}(x) = \max(0,x) + a \times \min(0,x)\,, \tag{4}$$

where a is a learnable parameter. Given a non-zero slop a, it can effectively alleviate the ‘dying ReLU’ problem.

The decoder architecture is based on the iterative up- and downsampling introduced by [50]. This design employs a series of alternating up-projection and down-projection operations. The transition between lower-resolution and higher-resolution feature spaces allows the network to effectively learn the intricate relationships between low-resolution and high-resolution features, leading to finer upsampling outputs. In this work, the iterative projection unit is implemented to facilitate the iterative up- and downsampling strategy. An iterative projection unit is made of four fundamental blocks depicted in Fig. 2. In the up block, the input low-resolution feature map L^{t-1} is upsampled to H_0^t, which is then downsampled back L_0^t. The difference between these two low-resolution feature maps is upsampled to H_1^t. Lastly, the sum of these two higher-resolution feature maps gives the final output H^t. The down block follows a similar process, focusing on dimensionality reduction; for the dense up block and dense down block, dense connections are introduced by concatenating all the previous low-resolution feature maps or high-resolution feature maps before applying the iterative up and downsampling operations.

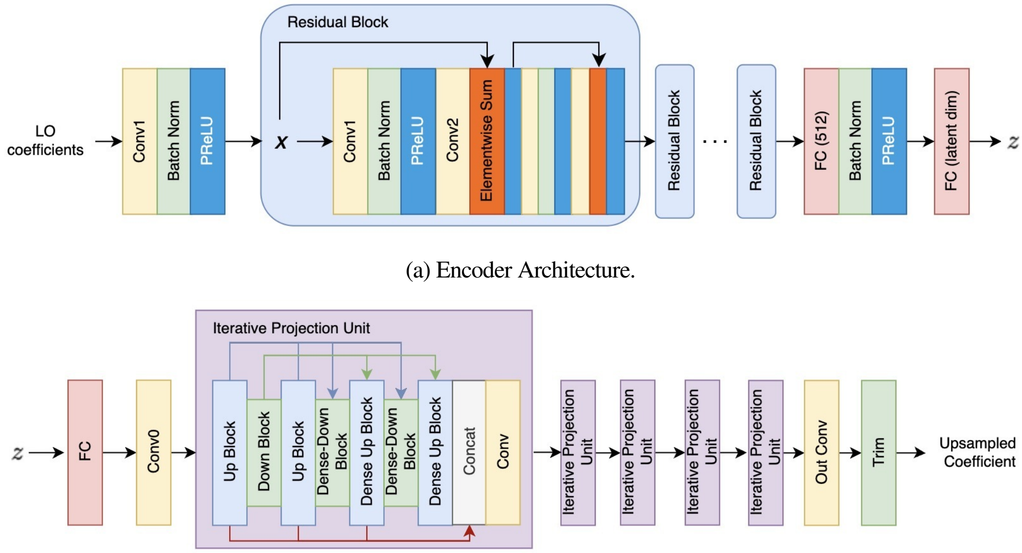
1.2.2 Discriminator network

The structure of the discriminator is illustrated in Fig. 3. The discriminator consists of nine convolutional layers, each followed by a batch normalization layer and a leaky rectified linear unit as the activation function. However, batch normalization is intentionally excluded from the first layer to prevent issues such as sample oscillation and model instability, as indicated in [51]. In order to create customized HRTF data for each individual, it is essential that the generated SH coefficients are diverse. To achieve this, the minibatch discrimination mechanism proposed by [52] is incorporated into the discriminator network. The process is as follows: for any sample $x_i \in \mathbb{R}^A$ in a batch, it is multiplied by a learnable matrix $T \in \mathbb{R}^{A\times B\times C}$, producing a feature map $M_i \in \mathbb{R}^{B\times C}$, where B stands for the number of features and C represents the dimensionality of each feature.

The diversity in the b-th feature between sample x_i and other samples is computed as $o(x_i)_b = \sum_{j=1}^{N} exp(-||M_{i,b} - M_{j,b}||_{L1}) \in \mathbb{R}$. Thus, each sample x_i has a corresponding vector $o(x_i) = [o(x_i)_1, o(x_i)_2, ..., o(x_i)_B] \in \mathbb{R}^B$, which contains diversity information. This vector $o(x_i)$ is then concatenated with the original input x_i to produce the output of the minibatch discrimination block. This mechanism allows the discriminator to consider the diversity of samples when assessing the authenticity. Lastly, two fully connected layers are employed to reduce the dimensionality of extracted features, and a sigmoid activation function is applied to the output for binary classification.

1.3 Loss Functions

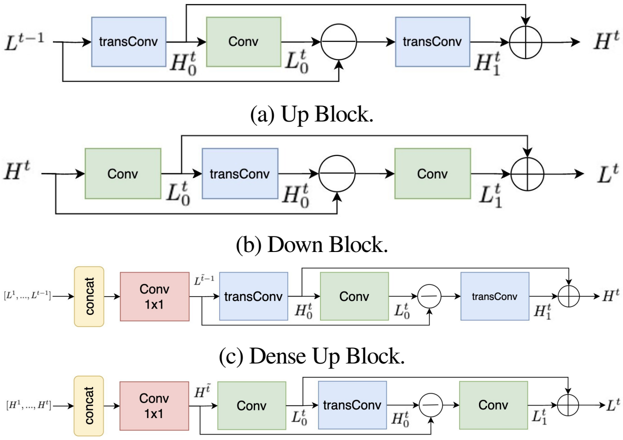
In an adversarial training framework, the network is typically optimised using adversarial. In this study, two additional loss functions are introduced: cosine similarity loss $\mathscr{L}_{\text{cos}}^{\text{G}}$ and content loss $\lambda\mathscr{L}_{\text{C}}^{\text{G}}$. Therefore the final loss function for the generator$\mathscr{L}^{\text{G}}$



(a) Encoder Architecture.

(b) Decoder Architecture.

Fig. 1: Autoencoder architecture taken from [47]. The blue and green arrows represent the dense connection for Dense Down Blocks and Dense Up Blocks, respectively. The red arrow indicates the concatenation of upsampled feature maps.



(a) Up Block.

(b) Down Block.

(c) Dense Up Block.

(d) Dense Down Block.

Fig. 2: Basic Blocks in Decoder taken from [47].

is defined as:

$$\mathscr{L}^{\mathrm{G}} = \mathscr{L}^{\mathrm{G}}_{\mathrm{cos}} + \lambda \mathscr{L}^{\mathrm{G}}_{\mathrm{C}} + \mathscr{L}^{\mathrm{G}}_{\mathrm{A}}. \tag{5}$$

1.3.1 Adversarial loss

The discriminator evaluates the authenticity of the generated samples and provides a score that reflects how realistic the upsampled coefficients appear. The adversarial loss for the autoencoder is defined using binary cross-entropy loss over M training samples as follows:

$$\mathscr{L}^{\mathrm{G}}_{\mathrm{A}} = -\frac{1}{M}\sum_{m=1}^{M}\log(1-D(G(\mathbf{c}^{m}_{\mathrm{L}}))), \tag{6}$$

where $G(\mathbf{c}^{m}_{\mathrm{L}})$ represents the upsampled SH coefficients and $D(G(\mathbf{c}^{m}_{\mathrm{L}}))$ denotes the score returned by the discriminator. The adversarial loss for the discriminator involves contributions from both real and synthetic samples:

$$\mathscr{L}^{\mathrm{D}} = -\frac{1}{M}\sum_{m=1}^{M}\left[\log D(\mathbf{c}^{m}_{\mathrm{H}}) + \log(1-D(G(\mathbf{c}^{m}_{\mathrm{L}})))\right], \tag{7}$$

where $\mathbf{c}^{m}_{\mathrm{H}}$ represents a sample of high-resolution SH coefficients.

1.3.2 Cosine similarity loss

To further guide the autoencoder in producing realistic SH coefficients, a cosine similarity loss is employed to measure the closeness between the generated coefficients and the target high-resolution coefficients. The similarity is computed for each frequency bin, and the average across all frequency bins defines the cosines similarity loss:

$$\mathscr{L}^{\mathrm{G}}_{\mathrm{cos}} = \sqrt{\frac{1}{W}\sum_{w=1}^{W}\left(1-\frac{\mathbf{c}^{f_w}_{\mathrm{G}}\cdot\mathbf{c}^{f_w}_{\mathrm{H}}}{\|\mathbf{c}^{f_w}_{G}\|\|\mathbf{c}^{f_w}_{H}\|}\right)^2}, \tag{8}$$

where $\mathbf{c}^{f_w}_{\mathrm{G}}$ and $\mathbf{c}^{f_w}_{\mathrm{H}}$ denote generated SH coefficients and target high-resolution coefficients given a specific frequency f_w respectively. W is the number of frequency bins in the HRTF data.

1.3.3 Content loss

The content loss proposed in [29] evaluates the discrepancy between two sets of HRTF data through the LSD metric and the interaural level difference (ILD) metric. These metrics are adopted in this work to effectively guide the autoencoder in generating meaningful coefficients, ultimately enabling the production of realistic HRTFs. The content loss is the sum of the LSD and ILD metrics:

$$\mathscr{L}^{\mathrm{G}}_{\mathrm{C}} = \mathrm{LSD} + \mathrm{ILD}. \tag{9}$$

The LSD metric compares the magnitude spectrum between the generated HRTF data, denoted as H_{G} and the target HRTF data H_{HR} given a specific frequency f_w at position x_n. The LSD loss can be mathematically written as:

$$\mathrm{LSD} = \frac{1}{N}\sum_{n=1}^{N}\sqrt{\frac{1}{W}\sum_{w=1}^{W}\left(20\log_{10}\frac{|H_{\mathrm{HR}}(f_w,x_n)|}{|H_{\mathrm{G}}(f_w,x_n)|}\right)^2}, \tag{10}$$

where N represents the total number of positions.

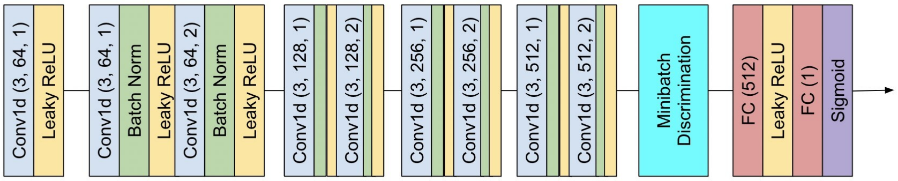


Fig. 3: Discriminator architecture.

The ILD measures the discrepancy between the magnitude of sound perceived by the left and right ears. The ILD loss is computed as the difference between the ILD of the generated HRTFs and of the target ones, expressed as:

$$\mathrm{ILD} = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{W}\sum_{w=1}^{W}\left|\left(20\log_{10}\frac{|H_{\mathrm{HR}}^{\mathrm{Left}}(f_w,x_n)|}{|H_{\mathrm{HR}}^{\mathrm{Right}}(f_w,x_n)|}\right) - \left(20\log_{10}\frac{|H_{\mathrm{G}}^{\mathrm{Left}}(f_w,x_n)|}{|H_{\mathrm{G}}^{\mathrm{Right}}(f_w,x_n)|}\right)\right|, \quad (11)$$

where $H^{\mathrm{Left}}(f_w,x_n)$ and $H^{\mathrm{Right}}(f_w,x_n)$ represent the left and right ear magnitude responses at frequency f_w and position x_n respectively.

2 EXPERIMENTAL SETUP

2.1 Data

The SONICOM HRTF dataset [14] was used for training and evaluating the proposed model. The dataset contains HRTF measurements from 203 subjects covering both left and right ears. For each HRTF set it includes HRIRs captured from 793 distinct positions distributed across a spherical surface with a radius of r = 1.5m. There are 72-row angles (azimuths) ranging from -180° to 180°, and 12 column angles (elevations) ranging from -45° to 90°. Notably, measurements around the horizontal plane are more densely distributed due to the smaller elevation interval. This design decision in HRTF measurement systems signifies an increased precision in human sound localization in this region. It is worth mentioning that an equal number of measurements were taken for each elevation. However, this arrangement might not be true for other HRTF datasets, such as the Acoustic Research Institute (ARI) HRTF database [53], where fewer measurements are available at high and low elevations.

Following the preprocessing outlined in Section 1.1, the resulting dimensions of the low-resolution HRTF, the chosen order of SH transformation, and the coefficient sizes are summarized in Table 1. Take note that the first row shows the size of the original full HRTF data while an order of 21 is selected to perform the SH transformation on high-resolution HRTF data, and this order is also used during postprocessing to inverse the generated SH coefficients back to the magnitude of HRTFs. The azimuth and elevation values in the first and second columns specify the number of angles covered in each respective direction. It is important to note that the number of azimuth angles chosen for each elevation does not need to be uniform. For instance, in the case of the third row with 19 points, the arrangement includes a total of four elevations (-30°, 0°, 30°, and 60°). In this configuration, 7 points are initially positioned at 7 unique azimuth angles for an elevation of 0°, while 4 distinct azimuth angles are distributed across the other three elevations. Fig. 4 shows the positions of the sources for 100 → 793, 19 → 793, 5 →793, 3 → 793.

Table 1: Downsampled HRTF size and corresponding SH transformation order and coefficients.

No. Azimuth	No. Elevation	No. Initial Points	SH Order	No. Coefficients
72	12	793	21	484
20	5	100	8	81
11	4	19	3	16
3	3	5	1	4
3	2	3	1	4

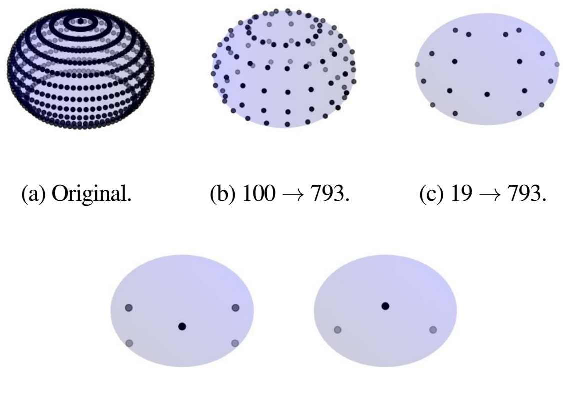
Theoretically, when utilizing all available data points for the SH transformation, the maximum order that can be used to represent the high resolution HRTF can be as high as 27 (note that $(27+1)^2 = 784 < 793$). This is because the 793 measurements are evenly distributed around the sphere except for missing measurements from below the subject (elevations below -45°).

The downsampled HRTFs, on the other hand, do not always contain measurements that are evenly distributed on the sphere. The consequence of this and of truncating the SH order causes a ‘spatial leakage’ effect and an increase in spatial aliasing. Therefore, the network needs to remove these distortions when upsampling the SH order to create the high-resolution HRTFs, as the target SH coefficients do not contain these artefacts. This, therefore, becomes a sort of denoising/enhancement problem as the network not only aims to upsample the SH order but also needs to remove any binaural artefacts caused by few and non-uniform HRTF measurements in the low-resolution HRTF.

To make this approach more computationally and memory efficient, we also reduce the SH order used to represent the high-resolution HRTFs to 21 instead of the maximum possible SH order of 27. This is possible because the low-order coefficients are much more important for the reconstruction than the higher order ones. Therefore, reducing the SH order to 21 will not substantially influence the quality of the final reconstructed HRTFs. Another added benefit of this order reduction is removing some of the issues surrounding the reconstruction of a high number of coefficients from the latent space, which could be exceedingly complex for the decoder. The discriminator is also easily able to distinguish between authentic high-order coefficients and upsampled coefficients. This phenomenon can potentially be attributed to the fact that even assuming each coefficient only deviates slightly from its ground truth value, these small errors might accumulate substantially as they pass through successive convolutional layers in the discriminator.

2.2 Model Training

It is important to note that in the SONICOM HRTF dataset, the left and right HRTFs for the same subject are stored separately. To facilitate the computation of interaural level differences during training, the left and right HRTFs are merged along the frequency dimension, effectively doubling the number of frequency bins.



(a) Original.

(b) 100 → 793.

(c) 19 → 793.

(d) 5 → 793.

(e) 3 → 793.

Fig. 4: The source positions for each sparsity level.

Table 2: Hyperparameter values for AE-GAN with different sparsity levels.

Hyperparameter	No. Initial Points			
	100	19	5	3
No. Epochs	200	200	200	200
Content Weight	0.02	0.02	0.01	0.01
Batch Size	4	6	8	8
LR - Generator	$1\times e^{-4}$	$1\times e^{-4}$	$2\times e^{-4}$	$2\times e^{-4}$
LR - Discriminator	$2\times e^{-5}$	$2\times e^{-5}$	$3\times e^{-5}$	$3\times e^{-5}$

Specifically, the first 128 frequency bins represent the left ear, while the subsequent 128 bins pertain to the right ear. This operation does not affect the number of points used for SH transformation. Among the 203 left-right pairs available, 162 are utilized for training, while the remaining 41 are reserved for evaluation.

During training, the AE-GAN model generates upsampled coefficients, which are multiplied by SHs to obtain high-resolution HRTFs. These upsampled HRTFs, along with the original high-resolution HRTFs, are employed to compute the content loss as described in Section 1.3.3.

The AE-GAN is trained separately for different sparsity levels, as detailed in Table 2. A content weight of 0.02 is applied for lower sparsity levels, while it is reduced to 0.01 for higher sparsity levels. The batch size ranges from 4 to 8, and the *Adam* optimizer [54] is used with hyperparameters $\beta_1 = 0.9$ and $\beta_2 = 0.9$. The generator's learning rate varies from $1e{-}4$ to $2e{-}4$, which is higher than the discriminator's learning rate of $2e{-}5$ to $3e{-}5$ to accelerate meaningful interpolation during initial stages. The discriminator is updated four times more frequently than the generator to provide accurate and constructive feedback.

The encoder input size depends on the input's sparsity levels, which adjusts the number of downsampling operations. These operations occur at most once in each residual block, with the stride of the first convolutional layer set to 2 and subsequent layers set to 1. For extreme cases, such as when only 4 coefficients are available, no downsampling is performed to retain sufficient information before encoding into the latent space. The detailed configuration is illustrated in Table 3.

Table 3: Residule block configuration for different input sizes.

No. Coefficient	Residual Block Configuration
25	[1, 2, 2, 2, 1]
81	[1, 2, 2, 2, 1]
16	[1, 2, 2, 1, 1]
4	[1, 1, 1, 1, 1]

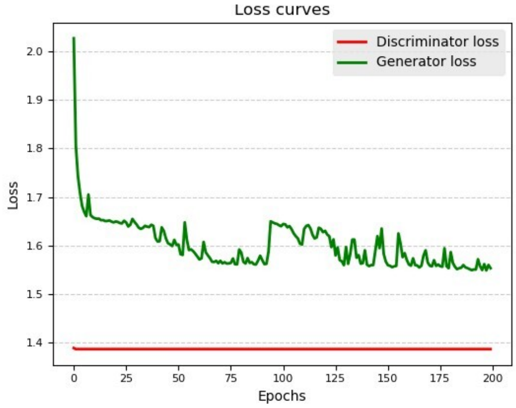


Fig. 5: Loss curves for discriminator and generator.

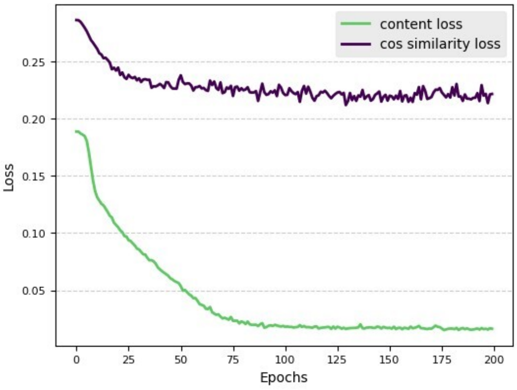


Fig. 6: Content loss and cosine similarity loss.

The training loss curves for both the generator and discriminator are illustrated in Fig. 5. The flat red line indicates that the discriminator converges quickly during training, suggesting it is effectively classifying real and generated SH coefficients. The generator loss starts above 2.0 and drops significantly at early steps since the upsampled SH coefficients deviate from the real samples by a large margin and stabilise around 1.55 after 160 epochs. The decreasing loss indicates that the generator is improving over time, producing samples that are closer to the ground truth targets.

The cosine similarity loss measures the distance between the reconstructed SH coefficients and the target coefficients, which is also an indicator of the generator's capability in upsampling coefficients. However, the ultimate goal is to produce realistic high-resolution HRTFs, and the quality of these HRTFs is assessed through the content loss. To examine how the improvement in coefficient similarity impacts the generated HRTFs' quality, the content loss curve and cosine similarity loss curve are plotted in Fig. 6. It can be seen that the content loss drops significantly as the cosine similarity decreases during the first 75 epochs. This trend shows a positive correlation between the quality of the generated coefficients and the resultant HRTFs, implying that enhancing the quality of coefficients leads to an improved final HRTF output.

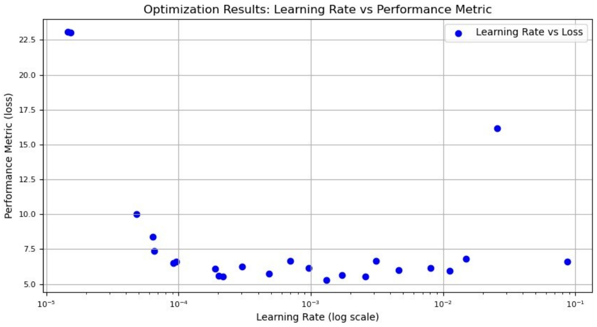


Fig. 7: Bayesian optimization of AE-GAN.

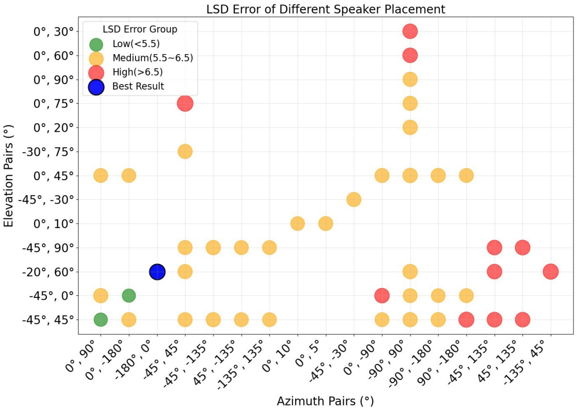


Fig. 8: LSD error for different speaker placement.

3 MODEL OPTIMIZATION

3.1 Bayesian Optimization

The learning rate is a critical parameter in neural network training, influencing both the efficiency of the optimization process and the model's final performance. Choosing an inappropriate learning rate can result in slow convergence, suboptimal performance, or even instability in training. To systematically identify the optimal learning rate, Bayesian optimization is employed to explore the hyperparameter space.

The search space for the learning rate is defined as a logarithmic uniform distribution ranging from 10^{-5} to 10^{-1}, evaluated over 200 training epochs. To efficiently navigate this space, the Tree-structured Parzen Estimator (TPE) algorithm is utilized [55]. Unlike traditional grid or random search methods, TPE adapts its exploration by modelling the search space probabilistically, focusing on promising regions based on prior evaluations. The evaluation criterion used in this optimization is the LSD metric, which provides a quantitative measure of the model's spectral distortion, reflecting its performance under varying learning rates.

The results, depicted in Fig. 7, offer significant insights. At extremely low learning rates ($< 10^{-4}$), the loss is higher, consistent with the expectation that such rates impede efficient weight updates and risk convergence to poor local minima. Conversely, excessively high learning rates destabilize training, causing fluctuations or divergence. Between these extremes lies a well-defined "sweet spot" where the learning rate minimizes the loss, achieving a balance between rapid convergence and training stability. The curve also exhibits a general decay trend, indicating progressive improvement in the model's performance as the learning rate approaches the optimal range.

Through these experiments, the optimal learning rate is identified as 0.065, achieving a mean LSD error of 5.032. This optimal value highlights the effectiveness of the TPE algorithm in uncovering hyperparameters that enhance model performance, and it underscores the importance of systematic optimization in neural network training.

3.2 Optimal Speaker Placement

The selection of initial points is crucial in determining the accuracy of upsampling results, particularly under sparse measurement scenarios. Each chosen point must encapsulate as much spatial and auditory information as possible to maximize the precision of the reconstruction process. For a full set of individual HRTF measurements, 793 points are typically used. These points are captured using strategically positioned speakers around the user, forming a complete auditory spatial dataset (as shown in Fig. 4a).

In scenarios requiring sparse measurements, such as with only four initial points, the placement of these initial points becomes even more critical. To evaluate and optimize the effectiveness of these points, the LSD metric is employed as a robust quantitative tool. The LSD metric measures the spectral fidelity of the reconstructed HRTFs, making it ideal for assessing the impact of initial point selection on the quality of upsampling in sparse setups.

The spatial arrangement of these initial points significantly influences the ability to reconstruct the auditory field accurately. Optimal placement ensures that the selected points capture essential auditory spatial cues, including interaural time differences (ITDs) and ILDs, which are key for reconstructing accurate HRTFs. Table 8 presents the results of the speaker placement selection process.

The speaker placement strategy prioritized leveraging spatial distribution to minimize LSD error and enhance HRTF upsampling accuracy. Initial experiments focused on key azimuthal points, such as -45° and 45°, to ensure critical left and right auditory cues were included. Subsequent testing revealed that while specific column points had minimal individual impact, configurations with uniformly distributed points consistently improved reconstruction accuracy. Further validation through diagonal configurations and evenly spaced azimuth points confirmed the advantage of symmetric and balanced spatial distribution.

The lowest LSD error (4.912) was achieved using a symmetric and evenly spaced configuration, with speakers placed at -180° and 0° azimuth and elevations of -20° and 60°. This result underscores the importance of symmetry and uniform spatial coverage in point placement. Such strategies ensure that the selected points provide optimal spatial cues for accurate HRTF reconstruction, particularly in scenarios where sparse data necessitates efficient use of limited information.

4 EXPERIMENTAL VALIDATION

4.1 Baselines

To evaluate the performance of the proposed approach, the AE-GAN model is tested on 41 subjects (not seen in training) from the SONICOM dataset and compared against three baselines: barycentric interpolation, SH interpolation, and non-individual HRTF selection.

The complete AE-GAN implementation and pre-processing code to reproduce these results are available at [56].

4.1.1 Barycentric interpolation

Barycentric interpolation is a robust technique well-suited for interpolating values within a simplex by leveraging weighted averages of the function's values at the vertices of the simplex. This method, which employs three barycentric coordinates as weights assigned to data points, enables the interpolation of unknown values within a set of known data points based on their surrounding values. Barycentric interpolation is utilized as a baseline to benchmark the performance of the proposed model.

4.1.2 SH interpolation

SH interpolation is a widely-used method for HRTF upsampling [57]. This technique projects HRTF data onto SHs, enabling a smooth and continuous representation across the spatial domain. By leveraging the mathematical properties of these spherical basis functions, SH interpolation accurately fills in missing data points, enhancing the resolution and accuracy of the upsampled HRTF. The corresponding SH orders for different sparsity levels are shown in Table 1.

4.1.3 Non-individual HRTF selection

Additionally, an alternative approach to modelling individualized HRTFs involves selecting the best-fitting HRTF from a database. Following the selection methodologies outlined in [29], two distinct sets of HRTFs are identified from the test set. These selections are based on their average LSD error, which quantifies their deviation from the other HRTFs in the training dataset. Selection-1 represents the subject whose HRTF yields the lowest average LSD error, suggesting that it is the most representative or generic among the dataset. Conversely, Selection-2 identifies the subject whose HRTF exhibits the highest average LSD error, indicating that it is the most distinctive or unique.

4.1.4 Two state-of-art methods

The proposed approach is also evaluated on two state-of-the-art methods that competed in the LAP challenge [46] and are available online. 1) SUpDEQ-MCA, a hybrid approach combining SUpDEq (Spatial Upsampling by Directional Equalization) [35], and MCA (Magnitude-Corrected and Time-Aligned Interpolation) [38], utilizes natural-neighbor interpolation for spatial upsampling, followed by magnitude correction and time alignment to refine the interpolated results. This method is particularly effective in reconstructing spatial audio data under low sparsity conditions due to its ability to balance spatial accuracy with temporal precision. 2) GEP-GAN (Gnomonic Equiangular Projection-GAN) [29], which employs a similar GAN-based structure to AE-GAN, demonstrates competitive performance, particularly at intermediate sparsity levels. It converts spherical HRIR data into a format for convolutional neural network processing using two steps: projection onto a 2D surface and interpolation onto an evenly spaced Cartesian grid.

4.2 LSD Metric Evaluation

The comparative analysis in Fig. 9 highlights the strengths and limitations of state-of-the-art methods, including SUpDEQ-MCA, GEP-GAN, and the proposed AE-GAN model, across varying sparsity levels [46]. The box plot visualizes the mean LSD error with standard deviation as an error measure. The median line represents the central tendency, while the box interquartile range and whiskers illustrate the variability in error. The standard deviation is used to generate synthetic samples, reflecting performance consistency across sparsity levels. Outliers are excluded to focus on the primary distribution. Note that LSD evaluation results of AE-GAN in this paper are differ slightly compare to previous work [47]due to enhancements in AE-GAN's architecture and optimization strategy.

Table 4: Mean LSD error (standard deviation) for AE-GAN across horizontal and vertical planes for different sparsity levels. The 'best' result of each sparsity level has been highlighted.

Method	Upsampling [No. initial → No. upsampled]			
	100 → 793	19 → 793	5 → 793	3 → 793
Horizontal plane	0.395 (0.41)	0.449 (0.45)	0.575 (0.53)	0.32(0.55)
Vertical plane	0.070 (0.11)	0.061 (0.21)	0.031 (0.13)	0.034 (0.19)

SUpDEQ-MCA demonstrates exceptional performance at low sparsity levels, achieving the lowest LSD error of 2.93 at 100 initial points. However, its performance deteriorates significantly under extreme sparsity, with LSD errors increasing to 6.67 at 3 initial points. In contrast, AE-GAN maintains competitive performance across all sparsity levels, outperforming SUpDEQ-MCA at higher sparsity levels with LSD errors of 4.59 and 4.76 at 5 and 3 initial points, respectively. This indicates AE-GAN's robustness in sparse data scenarios, where traditional methods and even state-of-art approaches falter. GEP-GAN achieves the best LSD error of 4.11 at 19 initial points. However, GEP-GAN's performance declines under extreme sparsity, with an LSD error of 5.21 at 3 initial points, underscoring AE-GAN's superior ability to handle sparse HRTF datasets despite architectural similarities.

Traditional interpolation methods, such as SH and Barycentric interpolation, exhibit substantial limitations as sparsity increases. While SH performs relatively well at lower sparsity, it suffers severe degradation at higher sparsity, with LSD errors escalating to 10.3 and 9.95 for 5 and 3 initial points, respectively. Similarly, Barycentric interpolation struggles under sparse conditions, with LSD errors rising to 8.33 and 8.55. This is because the barycentric interpolation and SH could not accurately estimate the value at the target position when the neighboring measured ones are far away from the desired location.

In contrast, the proposed AE-GAN has learned from low and high-order coefficient pairs, enabling it to reconstruct the spherical harmonics that closely represent the whole set of HRTF measurements. Therefore, irrespective of the sparsity levels applied and the separation between the measured points and the target position, the AE-GAN is able to predict the values decently. While Selection-1 slightly outperforms barycentric interpolation and SH at certain sparsity levels demonstrate modest improvements over traditional interpolation techniques, they are consistently outperformed by AE-GAN and other advanced methods, particularly in extremely sparse scenarios.

To further evaluate the spatial distribution of errors in HRTF reconstruction, Table 4 presents a comparative analysis of the LSD errors across horizontal and vertical planes for different sparsity levels. The horizontal plane consistently demonstrates higher LSD errors and larger standard deviations, indicating challenges in reconstructing azimuth-related spatial cues. The highest LSD error of 0.575 is observed at the sparsity level with 5 initial points, whereas the lowest error of 0.32 is achieved at the sparsity level with 3 initial points. This variability highlights the model's sensitivity to sparsity levels in reconstructing azimuthal features.

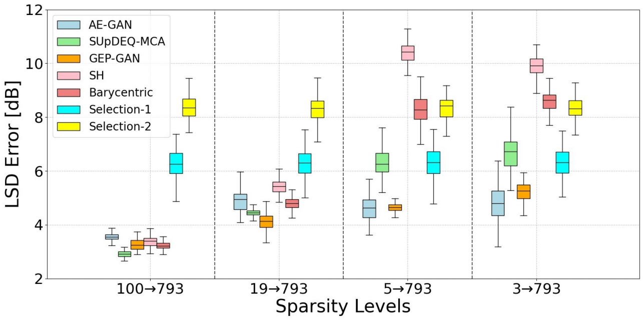


Fig. 9: Log-spectral distance (LSD) error across sparsity levels. Each box shows the median, interquartile range, and 95% confidence interval based on synthesized data from reported means and standard deviations.

Conversely, the vertical plane exhibits consistently low LSD errors with minimal variability, emphasizing the model's strength in capturing elevation-related spatial cues. The lowest error of 0.031 is observed at the sparsity level with 5 initial points, with similarly low values across other sparsity levels. This stability suggests that the model benefits from either sufficient training data coverage or inherent design optimizations that prioritize elevation-based features.

4.3 Perceptual Auditory Model Evaluation

Another metric under consideration is the evaluation of localization performance, which can be assessed by a Bayesian model, such as Barumerli2023 [58]. Evaluating localization performance is crucial as it relates to human perception. It's worth noting that even minor errors in LSD can have a significant impact on localization performance. In some cases, even substantial LSD errors may not affect localization performance to the same degree.

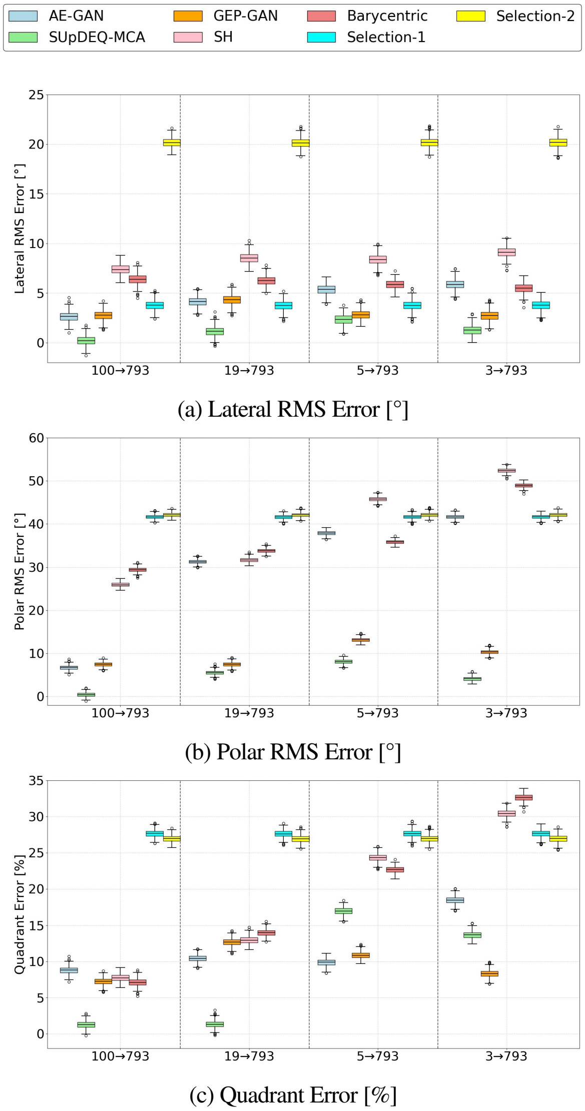
To conduct this evaluation, it is necessary to add the phase information into the aligned left and right HRTFs. This is achieved by using a minimum-phase approximation and a simple ITD model outlined in [29] [59]. The resultant complete HRTFs are passed into directional transfer functions (DTFs) [60], and the output features are subsequently input into the Barumerli2023 model for further analysis.

The 'Target' result is obtained by assessing the localization performance of the original high-resolution HRTF against itself, indicating the best performance that can be achieved. The evaluation of the proposed method and four baselines are carried out by comparing their results with the benchmark set by the 'Target' standard.

The results derived from the Barumerli2023 model are presented with a corresponding graphical representation illustrated in Fig. 10. Lateral root mean square (RMS) error, polar RMS error and quadrant error as described in [61,62], are mathematically defined for N localization trials, where each target source direction ϕ_i is paired with its response direction $\tilde{\phi}_i$, for $i=1,2,...,N$. A subset of local responses is defined as $\mathscr{A}=\{i:\mathrm{wrap}|\tilde{\phi}_i-\phi_i|<90^\circ\}$, and the metrics are calculated as follows:

$$\text{Lateral RMS Error}=\sqrt{\frac{\sum_{i\in\mathscr{A}}((\tilde{\phi}_i-\phi_i))^2}{|\mathscr{A}|}}, \tag{12}$$

$$\text{Polar RMS Error}=\sqrt{\frac{\sum_{i\in\mathscr{A}}(\mathrm{wrap}(\tilde{\phi}_i-\phi_i))^2}{|\mathscr{A}|}}, \tag{13}$$



(a) Lateral RMS Error [°]

(b) Polar RMS Error [°]

(c) Quadrant Error [%]

Fig. 10: Results from the simulated perceptual evaluation. Each box plot visualizes median, interquartile range, and 95% confidence intervals for different localization performance metrics.

$$\text{Quadrant Error}=\left(1-\frac{|\mathscr{A}|}{N}\right)\times 100. \tag{14}$$

The evaluation results in Fig. 10 reaffirm the superior performance of the AE-GAN model across various sparsity levels, demonstrating its robustness compared to baseline methods. For lateral RMS Error, AE-GAN achieves the best results for low sparsity levels, maintaining a competitive edge with errors of 2.63 and 4.12 at 100 and 19 initial points, respectively. GEP-GAN performs best at extreme sparsity, achieving the best performance at 5 and 3 initial points with errors of 2.83 and 2.75, indicating a strong capability for handling sparse datasets. This result indicates that with more initial points, the AE-GAN model benefits from a richer set of input data. It enables AE-GAN's latent space representation to accurately capture intricate spatial cues and project them into corresponding positions.

For polar RMS error, AE-GAN demonstrates consistent superiority in low sparsity levels, achieving the best error of 6.74 at 100 initial points. At higher sparsity, GEP-GAN achieves strong results, particularly in extreme sparsity conditions. Its

GAN-based structure, which is similar to AE-GAN, allows it to effectively handle sparse input data. This is demonstrated by its recorded errors of 13.17 and 10.35 at 5 and 3 initial points, respectively, further showcasing its adaptability in such scenarios. Traditional methods like barycentric interpolation and SH also struggle at higher sparsity, where their geometric assumptions just break apart.

In terms of quadrant error, AE-GAN shows strong adaptability across sparsity levels, achieving competitive errors of 10.41 and 9.83 at 19 and 5 initial points, respectively. While SH and Barycentric interpolation shows relatively better performance at low sparsity levels (errors of 7.75 and 7.12, respectively, at 100 initial points), their performance deteriorates significantly at higher sparsity, with errors exceeding 30.42 for SH and 32.65 for Barycentric at 3 initial points. GEP-GAN achieves notable results in this metric as well, with an error of 8.34 at 3 initial points, highlighting its robustness in sparse conditions. Interestingly, this observation highlights that performance on polar RMS error and lateral RMS error does not directly correlate with quadrant error, as improvements in the former metrics may not necessarily translate to better quadrant error performance and could, in some cases, contribute to worse results.

Selection-1 and Selection-2, while consistent across all sparsity levels, exhibit higher errors overall compared to AE-GAN and GEP-GAN. Selection-1 records errors of 3.78 for Lateral RMS Error, 41.67 for Polar RMS Error, and 27.67 for Quadrant Error, while Selection-2 records 8.33, 42.11, and 26.93, respectively. These results indicate limited adaptability of the selection-based approaches to varying sparsity levels.

5 DISCUSSION

While our study focused on the evaluation of GEP-GAN and SUpDEQ-MCA, it is essential to also consider the performance of other top-ranked methods submitted to the LAP Challenge. MERL-1 and MERL-2 [63], the winning submissions in Task 2, employed neural field architectures and retrieval-augmented generation to produce high-fidelity HRTFs across all sparsity levels. These methods showed strong consistency across evaluation metrics such as ITD, ILD, and LSD, particularly at higher sparsity levels such as 3→793 and 5→793. Their performance highlights the advantage of personalized model fine-tuning and data augmentation strategies when dealing with limited spatial information. In contrast, SYT-FSP-AE, which used a frequency-aware conditioned autoencoder, demonstrated moderate performance across LSD and perceptual metrics. While less competitive at extreme sparsity levels, it maintained robustness due to its generalizable structure that integrates position and frequency cues during reconstruction. This approach provides a middle ground between computational complexity and accuracy, particularly when adapting to unseen subjects or new datasets.

Interestingly, a key insight from the LAP Challenge was that LSD scores do not align well with perceptual localization performance, especially for machine learning-based methods. Some models that ranked highly on LSD, such as GEP-GAN, underperformed on polar and lateral RMS errors. In contrast, SUpDEQ-MCA [64, 65], although not winning overall, achieved the best results on polar RMS and quadrant error at high sparsity levels (e.g., 100 and 19 positions). This finding reinforces the importance of including perceptual auditory models in evaluation pipelines, beyond relying solely on spectral distortion metrics.

6 CONCLUSION

In this paper, it is shown that the proposed AE-GAN framework is capable of upsampling highly sparse HRTFs, especially in conditions where conventional interpolation methods prove inadequate. Instead of applying deep learning directly to the unevenly distributed HRTF data, the proposed approach transformed the HRTF data in the frequency domain into the SH domain. This transformation encoded the original spatial and frequency information into a set of SH functions and their associated coefficients, providing a structured and compact representation. The LSD evaluation highlights that the proposed deep learning model outperforms barycentric interpolation, particularly in extreme sparsity scenarios involving fewer than 12 measurements. These results underscore AE-GAN's robustness and effectiveness in addressing challenges posed when only sparse measurements are available.

In future work, alternative deep-learning models for SH coefficient extrapolation could be explored. Improving the weighting of harmonics could lead to more accurate representations of the original HRTFs, offering further improvements in reconstruction fidelity. The process of reconstructing high-order coefficients from low-order coefficients aligns well with sequence-to-sequence prediction frameworks. Another avenue to explore would be more advanced neural architectures, such as recurrent neural networks (RNNs) [66] and transformer models [67], as these models excel at processing variable-length inputs, making them particularly suitable for scenarios where the SH order varies based on the number of HRTF measurements. Such future studies could pave the way for more versatile and efficient HRTF upsampling techniques, further advancing the field.

7 ACKNOWLEDGMENT

This study was supported by the SONICOM project (www.sonicom.eu) funded by the European Union's Horizon 2020 research and innovation program under grant agreement 101017743.

8 REFERENCES

[1] D. Vickers, M. Salorio-Corbetto, S. Driver, C. Rocca, Y. Levtov, K. Sum, *et al.*, "Involving Children and Teenagers with Bilateral Cochlear Implants in the Design of the BEARS (Both EARS) Virtual Reality Training Suite Improves Personalization," *Front. Digit. Health*, vol. 3 (2021 Nov.).

[2] T. Lokki, H. Nironen, S. Vesa, L. Savioja, A. Härmä, and M. Karjalainen, "Application Scenarios of Wearable and Mobile Augmented Reality Audio," presented at the *Proc. Audio Eng. Soc. (AES) Conv.* (2004 May).

[3] F. L. Wightman and D. J. Kistler, "Headphone Simulation of Free-Field Listening. I: Stimulus Synthesis," *J. Acoust. Soc. Am.*, vol. 85, no. 2 (1989 Feb.), doi:10.1121/1.397557.

[4] V. Bruschi, L. Grossi, N. A. Dourou, A. Quattrini, A. Vancheri, T. Leidi, *et al.*, "A Review on Head-Related Transfer Function Generation for Spatial Audio," *Applied Sciences*, vol. 14, no. 23, p. 11242 (2024).

[5] J. Blauert, *Spatial hearing: the psychophysics of human sound localization* (MIT Press, Cambridge, Mass, 1983).

[6] P. Stitt, L. Picinali, and B. F. G. Katz, "Auditory Accommodation to Poorly Matched Non-Individual Spectral Localization Cues through Active Learning," *Scientific Reports*, vol. 9, no. 1, pp. 1063:1–14 (2019 Jan.), doi:10.1038/s41598-018-37873-0.

[7] E. M. Wenzel, M. Arruda, D. J. Kistler, and F. L. Wightman, "Localization Using Nonindividualized Head-Related Transfer Functions," *J. Acoust. Soc. Am.*, vol. 94, no. 1, pp. 111–123 (1993 Jul.), doi:10.1121/1.407089.

[8] H. Møller, M. F. Sørensen, C. B. Jensen, and D. Hammershøi, "Binaural Technique: Do We Need Individual Recordings?" *J. Audio Eng. Soc. (AES)*, vol. 44, no. 6, pp. 451–469 (1996 Jun.).

[9] Y. Kahana and P. A. Nelson, "Numerical Modelling of the Spatial Acoustic Response of the Human Pinna," *J. of Sound and Vibration*, vol. 292, no. 1, pp. 148–178 (2006 Apr.), doi:10.1016/j.jsv.2005.07.048.

[10] L. S. R. Simon, N. Zacharov, and B. F. G. Katz, "Perceptual Attributes for the Comparison of Head-Related Transfer Functions," *J. Acoust. Soc. Am.*, vol. 140, no. 5, pp. 3623–3632 (2016 Nov.), doi:10.1121/1.4966115.

[11] S. Werner, F. Klein, T. Mayenfels, and K. Brandenburg, "A Summary on Acoustic Room Divergence and Its Effect on Externalization of Auditory Events," presented at the *Proc. Int. Conf. on Quality of Multimedia Experience (QoMEX)*, pp. 1–6 (2016 Jun.), doi:10.1109/QoMEX.2016.7498973.

[12] I. Engel, D. L. Alon, P. W. Robinson, and R. Mehra, "The Effect of Generic Headphone Compensation on Binaural Renderings," presented at the *Proc. Audio Eng. Soc. (AES) Int. Conf. on Immersive and Interactive Audio* (2019 Mar.).

[13] M. Cuevas-Rodriguez, D. Gonzalez-Toledo, A. Reyes-Lecuona, and L. Picinali, "Impact of Non-Individualised Head Related Transfer Functions on Speech-in-Noise Performances within a Synthesised Virtual Environment," *J. Acoust. Soc. Am.*, vol. 149, no. 4, pp. 2573–2586 (2021 Apr.).

[14] I. Engel, R. Daugintis, T. Vicente, A. O. T. Hogg, J. Pauwels, A. J. Tournier, *et al.*, "The SONICOM HRTF Dataset," *J. Audio Eng. Soc. (AES)* (2023 Jun.), doi: 10.17743/jaes.2022.0066.

[15] H. Ziegelwanger, W. Kreuzer, and P. Majdak, "Mesh2HRTF: An Open-Source Software Package for the Numerical Calculation of Head-Related Transfer Functions," presented at the *Proc. Int. Cong. on Sound and Vibration (ICSV)*, pp. 1–8 (2015 Jul.), doi:10.13140/RG.2.1.1707.1128.

[16] B. F. G. Katz, "Boundary Element Method Calculation of Individual Head-Related Transfer Function. I. Rigid Model Calculation," *J. Acoust. Soc. Am.*, vol. 110, no. 5, pp. 2440–2448 (2001 Nov.), doi:10.1121/1.1412440.

[17] P. Stitt and B. F. Katz, "Sensitivity Analysis of Pinna Morphology on Head-Related Transfer Functions Simulated via a Parametric Pinna Model," *J. Acoust. Soc. Am.*, vol. 149, no. 4, pp. 2559–2572 (2021 Apr.), doi:10.1121/10.0004128.

[18] M. Geronazzo, E. Peruch, F. Prandoni, and F. Avanzini, "Applying a Single-Notch Metric to Image-Guided Head-Related Transfer Function Selection for Improved Vertical Localization," *J. Audio Eng. Soc. (AES)*, vol. 67, no. 6, pp. 414–428 (2019 Jun.).

[19] DYN. Zotkin, J. Hwang, R. Duraiswaini, and L. S. Davis, "HRTF Personalization Using Anthropometric Measurements," presented at the *Proc. IEEE Workshop on Appl. of Signal Process. to Audio and Acoust. (WASPAA)*, pp. 157–160 (2003 Oct.), doi:10.1109/ASPAA.2003.1285855.

[20] B. F. Katz and G. Parseihian, "Perceptually Based Head-Related Transfer Function Database Optimization," *J. Acoust. Soc. Am.*, vol. 131, no. 2, pp. EL99–EL105 (2012 Feb.).

[21] C. Kim, V. Lim, and L. Picinali, "Investigation into Consistency of Subjective and Objective Perceptual Selection of Non-Individual Head-Related Transfer Functions," *J. Audio Eng. Soc. (AES)*, vol. 68, no. 11, pp. 819–831 (2020 Dec.).

[22] F. Zagala, M. Noisternig, and B. F. Katz, "Comparison of Direct and Indirect Perceptual Head-Related Transfer Function Selection Methods," *J. Acoust. Soc. Am.*, vol. 147, no. 5, pp. 3376–3389 (2020 May).

[23] L. Picinali and B. F. G. Katz, "System-to-User and User-to-System Adaptations in Binaural Audio," in Sonic Interactions in Virtual Environments," in M. Geronazzo and S. Serafin (Eds.), *Sonic Interactions in Virtual Environments*, pp. 121–144 (Springer, 2022 Oct.).

[24] T. Carpentier, H. Bahu, M. Noisternig, and O. Warusfel, "Measurement of a Head-Related Transfer Function Database with High Spatial Resolution," presented at the *Proc. EAA Forum Acusticum, Eur. Congress on Acoust.* (2014 Sep.).

[25] A. Farina, "Simultaneous Measurement of Impulse Response and Distortion with a Swept-Sine Technique," presented at the *Proc. Audio Eng. Soc. (AES) Conv.*, pp. 1–23 (2000 Feb.).

[26] D. N. Zotkin, R. Duraiswami, E. Grassi, and N. A. Gumerov, "Fast Head-Related Transfer Function Measurement via Reciprocity," *J. Acoust. Soc. Am.*, vol. 120, no. 4, pp. 2202–2215 (2006 Oct.), doi:10.1121/1.2207578.

[27] J.-G. Richter, G. Behler, and J. Fels, "Evaluation of a Fast HRTF Measurement System," presented at the *Proc. Audio Eng. Soc. (AES) Conv.*, vol. 140, p. 9498 (2016 May).

[28] X.-L. Zhong and B.-S. Xie, *Head-Related Transfer Functions and Virtual Auditory Display* (InTech, Plantation, FL, USA, 2014 Mar.), doi:10.5772/56907.

[29] A. O. T. Hogg, J. Mads, H. Liu, I. Squires, S. J. Cooper, and L. Picinali, "HRTF Upsampling with a Generative Adversarial Network Using a Gnomonic Equiangular Projection," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 32, pp. 2085 – 2099 (2024 Mar.).

[30] K. Hartung, J. Braasch, and S. J. Sterbing, "Comparison of Different Methods for the Interpolation of Head-Related Transfer Functions," presented at the *Proc. Audio Eng. Soc. (AES) Conf. on Spatial Sound Reproduction* (1999 Mar.).

[31] D. Poirier-Quinot and B. F. G. Katz, "The Anaglyph Binaural Audio Engine," presented at the *Proc. Audio Eng. Soc. (AES) Conv.*, 144 (2018 May).

[32] M. Cuevas-Rodríguez, L. Picinali, D. González-Toledo, C. Garre, E. de la Rubia-Cuestas, L. Molina-Tanco, *et al.*, "3D Tune-in Toolkit: An Open-Source Library for Real-Time Binaural Spatialisation," *PLOS ONE*, vol. 14, no. 3, p. e0211899 (2019 Mar.).

[33] H. Gamper, "Head-Related Transfer Function Interpolation in Azimuth, Elevation, and Distance," *J. Acoust. Soc. Am.*, vol. 134, no. 6, pp. EL547–EL553 (2013 Dec.), doi:10.1121/1.4828983.

[34] M. J. Evans, J. A. S. Angus, and A. I. Tew, "Analyzing Head-Related Transfer Function Measurements Using Surface Spherical Harmonics," *J. Acoust. Soc. Am.*, vol. 104, no. 4, pp. 2400–2411 (1998 Jun.).

[35] C. Pörschmann, J. M. Arend, and F. Brinkmann, "Directional Equalization of Sparse Head-Related Transfer Function Sets for Spatial Upsampling," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 27, no. 6, pp. 1060–1071 (2019 Jun.), doi:10.1109/TASLP.2019.2908057.

[36] J. M. Arend, F. Brinkmann, and C. Pörschmann, "Assessing Spherical Harmonics Interpolation of Time-Aligned Head-Related Transfer Functions," *J. Audio Eng. Soc. (AES)*, vol. 69, no. 1/2, pp. 104–117 (2021 Feb.).

[37] I. Engel, D. F. M. Goodman, and L. Picinali, "Assessing HRTF Preprocessing Methods for Ambisonics Rendering through Perceptual Models," *Acta Acust.*, vol. 6, p. 4 (2022 Jan.), doi:10.1051/aacus/2021055.

[38] J. M. Arend, C. Pörschmann, S. Weinzierl, and F. Brinkmann, "Magnitude-Corrected and Time-Aligned Interpolation of Head-Related Transfer Functions," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 31, pp. 3783–3799 (2023 Sep.), doi:10.1109/TASLP.2023.3313908.

[39] T.-Y. Chen, T.-H. Kuo, and T.-S. Chi, "Autoencoding HRTFs for DNN Based HRTF Personalization Using Anthropometric Features," presented at the *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, pp. 271–275 (2019 May), doi:10.1109/ICASSP.2019.8683814.

[40] D. Yao, J. Zhao, L. Cheng, J. Li, X. Li, X. Guo, *et al.*, "An Individualization Approach for Head-Related Transfer Function in Arbitrary Directions Based on Deep Learning," *JASA Express lett. (JASA-EL)*, vol. 2, no. 6, p. 064401 (2022 Jun.), doi:10.1121/10.0011575.

[41] Y. Ito, T. Nakamura, S. Koyama, and H. Saruwatari, "Head-Related Transfer Function Interpolation from Spatially Sparse Measurements Using Autoencoder with Source Position Conditioning," presented at the *Proc. Int. Workshop on Acoust. Signal Enhancement (IWAENC)*, pp. 1–5 (2022 Sep.), doi:10.1109/IWAENC53105.2022.9914751.

[42] R. Duraiswami, D. N. Zotkin, and N. A. Gumerov, "Interpolation and Range Extrapolation of HRTFs," presented at the *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, vol. 4, pp. iv–45 (2004 May).

[43] G. Kestler, S. Yadegari, and D. Nahamoo, "Head Related Impulse Response Interpolation and Extrapolation Using Deep Belief Networks," presented at the *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, pp. 266–270 (2019 May), doi:10.1109/ICASSP.2019.8683570.

[44] Z. Jiang, J. Sang, C. Zheng, A. Li, and X. Li, "Modeling Individual Head-Related Transfer Functions from Sparse Measurements Using a Convolutional Neural Network," *J. Acoust. Soc. Am.*, vol. 153, no. 1, pp. 248–259 (2023 Jan.), doi:10.1121/10.0016854.

[45] X. Hu, J. Li, L. Picinali, and A. O. Hogg, "A Machine Learning Approach for Denoising and Upsampling HRTFs," *arXiv preprint arXiv:2504.17586* (2025).

[46] M. Geronazzo, L. Picinali, A. Hogg, R. Barumerli, K. Poole, R. Daugintis, *et al.*, "Technical Report: SONICOM / IEEE Listener Acoustic Personalisation (LAP) Challenge - 2024," (2024 Nov.), doi:10.36227/techrxiv.173153187.72930965/v1.

[47] X. Hu, J. Li, L. Picinali, and A. O. Hogg, "HRTF spatial upsampling in the spherical harmonics domain employing a generative adversarial network," *International Conference on Digital Audio Effects (DAFx)* (2024).

[48] A. Hogg, H. Liu, M. Jenkins, and L. Picinali, "Exploring the Impact of Transfer Learning on GAN-based HRTF Upsampling," presented at the *Proc. EAA Forum Acusticum, Eur. Congress on Acoust.* (2023 Sep.).

[49] E. Thuillier, C. Jin, and V. Välimäki, "HRTF Interpolation using a Spherical Neural Process Meta-Learner," *IEEE/ACM Trans. Audio, Speech, Language Process.* (2024).

[50] M. Haris, G. Shakhnarovich, and N. Ukita, "Deep back-projection networks for super-resolution," presented at the *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 1664–1673 (2018).

[51] A. Radford, L. Metz, and S. Chintala, "Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks," presented at the *Proc. Int. Joint Conf. on Learning Representations (ICLR)* (2015).

[52] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training gans," *Proc. Neural Inform. Process. Conf*, vol. 29 (2016).

[53] P. Majdak, "ARI HRTF Database," (2022 Jun.), URL `http://www.kfs.oeaw.ac.at/hrtf`.

[54] D. P. Kingma and L. J. Ba, "Adam: A Method for Stochastic Optimization," presented at the *Proc. Int. Joint Conf. on Learning Representations (ICLR)* (2015 May).

[55] Y. Ozaki, Y. Tanigaki, S. Watanabe, M. Nomura, and M. Onishi, "Multiobjective tree-structured Parzen estimator," *J. Artificial Intelligence Res.*, vol. 73, pp. 1209–1250 (2022).

[56] X. Hu, J. Li, L. Picinali, and A. O. Hogg (2025), URL `https://github.com/GeorgeHux/HRTF-Spatial-Upsampling-using-an-AE-GAN`.

[57] J. M. Arend, F. Brinkmann, and C. Pörschmann, "Assessing spherical harmonics interpolation of time-aligned head-related transfer functions," *J. Audio Eng. Soc. (AES)*, vol. 69, no. 1/2, pp. 104–117 (2021).

[58] R. Barumerli, P. Majdak, M. Geronazzo, D. Meijer, F. Avanzini, and R. Baumgartner, "A Bayesian model for human directional localization of broadband static sound sources," *Acta Acust.*, vol. 7, p. 12 (2023).

[59] A. Andreopoulou and B. F. Katz, "Perceptual impact on localization quality evaluations of common pre-processing for non-individual head-related transfer functions," *J. Audio Eng. Soc. (AES)*, vol. 70, no. 5, pp. 340–354 (2022).

[60] J. C. Middlebrooks, "Individual differences in external-ear transfer functions reduced by scaling in frequency," *J. Acoust. Soc. Am.*, vol. 106, no. 3, pp. 1480–1492 (1999).

[61] J. C. Middlebrooks, "Virtual Localization Improved by Scaling Nonindividualized External-Ear Transfer Functions in Frequency," *J. Acoust. Soc. Am.*, vol. 106, no. 3 Pt 1, pp. 1493–1510 (1999 Sep.).

[62] R. Baumgartner, P. Majdak, and B. Laback, "Modeling Sound-Source Localization in Sagittal Planes for Human Listeners," *J. Acoust. Soc. Am.*, vol. 136, no. 2, pp. 791–802 (2015 Sep.), doi:10.1121/1.4887447.

[63] M. Delcroix, K. Zmolikova, T. Ochiai, K. Kinoshita, S. Araki, and T. Nakatani, "Compact network for speakerbeam target speaker extraction," presented at the *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6965–6969 (2019).

[64] C. Pörschmann, J. M. Arend, and F. Brinkmann, "Directional equalization of sparse head-related transfer function sets for spatial upsampling," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 6, pp. 1060–1071 (2019).

[65] J. M. Arend, C. Pörschmann, S. Weinzierl, and F. Brinkmann, "Magnitude-corrected and time-aligned interpolation of head-related transfer functions," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 3783–3799 (2023).

[66] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, pp. 533–536 (1986).

[67] A. Vaswani, "Attention is all you need," *Proc. Neural Inform. Process. Conf* (2017).

THE AUTHORS

Xuyi Hu

Jian Li

Lorenzo Picinali

Aidan O. T. Hogg

Xuyi Hu is a PhD student at Imperial College London. He studied Electrical and Electronic Engineering at University College London for his undergraduate degree and then completed a Master's in Artificial Intelligence at Imperial College London. His research focuses on using machine learning to improve HRTF (head-related transfer function) upsampling and to measure personalised HRTFs.

●

Jian Li is currently an employee at Huawei, specializing in the training of large language models. He holds a Bachelor's degree in Electrical and Electronic Engineering from Nanyang Technological University (NTU) and a Master's degree in Artificial Intelligence from Imperial College London.

●

Lorenzo Picinali is a Professor in Spatial Acoustics and Immersive Audio and the lead of the Audio Experience Design at Imperial College London. In the past few years, he worked in Italy, France, and the United Kingdom on projects dealing with 3D binaural sound rendering, interactive applications for visually and hearing impaired individuals, audiology and hearing aid technology, audio and haptic interaction, and more in general, acoustical virtual and augmented reality. More information about the projects in which Lorenzo is involved can be found here: `https://www.axdesign.co.uk/.`

●

Aidan is a Lecturer in Computer Science and the co-lead for the Virtual, Immersive, Augmented and Binaural Audio Lab (VIABAL) in the Centre for Digital Music (C4DM) at Queen Mary University of London. He received an M.Eng. degree in electronic and information engineering and a PhD degree from Imperial College London in 2017 and 2022, respectively. His current research focuses on using deep learning to capture head-related transfer functions and, more generally, spatial acoustics and immersive audio. Other research interests include speaker diarization and statistical signal processing for audio applications. More information about current research projects can be found here: `https://aidanhogg.uk/`