

# HRTFformer: A Spatially-Aware Transformer for Personalized HRTF Upsampling in Immersive Audio Rendering

Xuyi Hu<sup>\*</sup> , Member, IEEE, Jian Li<sup>\*</sup> , Shaojie Zhang , Stefan Goetz , Member, IEEE, Lorenzo Picinali , Ozgur B. Akan , Fellow, IEEE, and Aidan O. T. Hogg

**Abstract**—Personalized Head-Related Transfer Functions (HRTFs) are starting to be introduced in many commercial immersive audio applications and are crucial for realistic spatial audio rendering. However, one of the main hesitations regarding their introduction is that creating personalized HRTFs is impractical at scale due to the complexities of the HRTF measurement process. To mitigate this drawback, HRTF spatial upsampling has been proposed with the aim of reducing measurements required. While prior work has seen success with different machine learning (ML) approaches, these models often struggle with long-range spatial consistency and generalization at high upsampling factors. In this paper, we propose a novel transformer-based architecture for HRTF upsampling, leveraging the attention mechanism to better capture spatial correlations across the HRTF sphere. Working in the spherical harmonic (SH) domain, our model learns to reconstruct high-resolution HRTFs from sparse input measurements with significantly improved accuracy. To enhance spatial coherence, we introduce a neighbor dissimilarity loss that promotes magnitude smoothness, yielding more realistic upsampling. We evaluate our method using both perceptual localization models and objective spectral distortion metrics. Experiments show that our model surpasses leading methods by a substantial margin in generating realistic, high-fidelity HRTFs.

**Index Terms**—immersive audio, head-related transfer function, transformer, upsampling, interpolation

## I. INTRODUCTION

Immersive audio often plays a vital role in applications such as virtual reality (VR) [1], [2], augmented reality (AR) [3], gaming [4], [5], and even therapeutic contexts [6] where it aims to recreate realistic spatial soundscapes that align with human auditory perception. Human spatial hearing relies on interaural and monaural localization cues. Interaural cues are typically categorized as interaural time differences (ITDs), which dominate at low frequencies, and interaural level differences (ILDs), which dominate at mid–high frequencies [7]. In the regions often termed as the ‘cone of confusion’ [8], where different source locations yield similar ITDs and ILDs, the auditory system exploits monaural spectral cues shaped by the pinnae. The centre frequency, depth, and placement of these pinna-induced spectral notches provide crucial information for elevation and for resolving front–back ambiguity. As expected, these spectral cues, together with ITDs and ILDs that depend on the listener’s head-and-torso morphology, are highly unique to each listener. These cues can all be captured by a person’s

This study was made possible by support from SONICOM ([www.sonicom.eu](http://www.sonicom.eu)), a project that has received funding from the European Union’s Horizon 2020 research and innovation program under grant agreement No. 101017743.

\* Equal contribution

Head-Related Transfer Function (HRTF), which describes how an individual’s anatomy filters sound from different directions before it reaches the eardrums [9]–[11].

It is well known that using non-individualized HRTFs, which are not personally tailored to a listener, can significantly compromise spatial audio performance. For example, studies have shown that generic HRTFs can often lead to impaired sound source localization, as the accurate spectral cues that are needed for spatial perception are strongly influenced by individual anatomical features, particularly the shape of the listeners’ pinnae [12], [13]. In addition to causing localization errors, non-individualized HRTFs have also been shown to negatively impact perceptual qualities such as externalization, immersion, timbral coloration, realism, and spatial depth [14]–[16]. Furthermore, the use of poorly matched HRTFs can reduce a listener’s ability to segregate and understand speech in complex auditory scenes, including multi-talker environments or in the cocktail party scenario [17]–[19]. These drawbacks highlight the need and importance of personalization of HRTFs to be able to deliver accurate and immersive auditory experiences [20]–[23].

In terms of HRTF personalization, various methods have been proposed, including direct acoustic measurements [24], 3D surface scanning [25], [26], anthropometry-based models [27]–[29], and selection from databases of measured HRTFs [30]. Among these, taking a direct acoustic measurement still remains the ‘gold standard’, as it is able to capture the listener-specific filtering effects precisely using in-ear microphones and controlled speaker arrays [31]–[33]. However, this approach is time-consuming, requires specialized equipment, and must be conducted in a noise-free environment, limiting its scalability and motivating the development of more practical alternatives.

To alleviate some of the downsides and difficulties with taking a direct acoustic measurement of a spatially dense HRTF, HRTF spatial upsampling has emerged as a promising alternative approach [34], [35]. It aims to reconstruct high-resolution HRTFs from a sparse set of acoustic measurements, significantly reducing the number of required sampling points. By leveraging spatial correlations and the underlying structure in HRTF data, upsampling methods enable efficient personalization while reducing measurement time and hardware needs.

HRTF spatial upsampling techniques are commonly divided into algorithmic and learning-based approaches. Algorithmic approaches estimate HRTFs at new source positions through interpolation, typically by combining existing measurements or basis functions derived from them [36], [37]. However, their performance degrades significantly in conditions when only a sparse measurement is available, as they rely heavily on dense

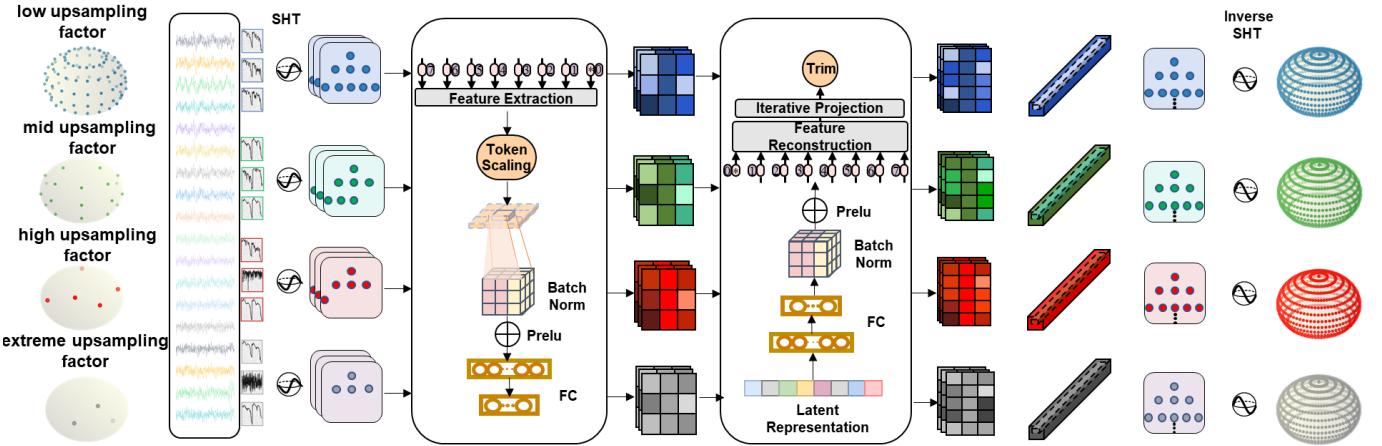


Fig. 1: The HRTF upsampling workflow of HRTFformer: the low-resolution HRTFs are first transformed into the spherical harmonics domain, which provides a compact and physically meaningful representation of directional acoustic information. The resulting SH coefficients serve as the model’s input. The encoder extracts global and local spatial features from the input and compresses them into the latent representation, which the decoder then uses to extrapolate and interpolate higher-degree SH coefficients. These are finally converted back into high-resolution HRTFs. This process is applied under four different sparsity conditions to evaluate the model’s robustness.

and evenly distributed input data. These approaches are also based on fixed mathematical assumptions, such as smoothness or symmetry, which may not accurately capture the complex, subject-specific variations in real HRTFs. These limitations have motivated the development of machine learning (ML) approaches - including supervised learning using morphological features [38], [39], neural networks trained on spatial HRTF data [40]–[42], and generative models that synthesize personalized HRTFs from sparse measurements [43]–[45]. Such approaches demonstrate better generalization and reconstruction quality, especially under sparse sampling conditions. However, they typically require large amounts of high-resolution training data and may struggle to generalize to unseen listeners.

This paper introduces HRTFformer (shown in Fig. 1), a spatially-aware transformer designed for HRTF upsampling. HRTFformer aims to overcome one of the main challenges that current ML methods face, which is that although they often achieve low log-spectral distortion (LSD), this frequently does not correlate well with perceptual performance (which is usually overlooked). It was found in [46] that the most state-of-the-art methods under high sparsity tend to generate general (average) HRTFs, which contradicts the goal of individual customization. To address this issue, our HRTFformer approach adopts an attention mechanism in transformers and introduces a neighbor dissimilarity loss, which promotes spatial continuity in the magnitude spectrum. It is shown in this work that HRTFformer is able to create more realistic and personalized HRTFs when compared to the latest state-of-the-art methods, especially in extremely sparse scenarios.

In summary, the contributions of this paper are as follows:

- 1) We propose a transformer-based architecture tailored for HRTF upsampling, named HRTFformer, which effectively captures global dependencies between sound energy distribution patterns, overcoming the limitations of performance degradation with spatially sparse data.
- 2) We introduce a novel neighbor similarity loss to enhance spatial consistency by respecting the natural variation in adjacent directions, thereby improving the realism and personalization of the reconstructed HRTFs.
- 3) We evaluate HRTFformer on both sparse and dense measurements, showing robust reconstruction performance.
- 4) We conduct comprehensive evaluations and demonstrate that HRTFformer achieves state-of-the-art results in both spectral evaluation and perceptual localization accuracy.

## II. RELATED WORK

It is common to separate existing methods into two categories: those that are Algorithmic-based and those that are Learning-based or data-driven.

### A. Algorithmic-based Approaches

Among algorithmic methods, barycentric interpolation [47], [48] and spherical harmonics (SH) interpolation [49], [50] are widely used. Barycentric interpolation estimates missing HRTFs by computing a weighted average of the three nearest neighbors, performing well when measurements are densely sampled (e.g., every 10–15°). However, its accuracy declines with sparser inputs (e.g., 30–40° spacing) due to increased distance between reference points. Similarly, SH interpolation represents the HRTF as a weighted sum of spatially continuous basis functions, the SHs. The SH coefficients (i.e. the contribution of each SH) are estimated by fitting this expansion to the measurements on the HRTF sphere (typically, least squares, sometimes weighted). When the number of samples is small relative to the chosen SH order, the fit becomes ill-conditioned: high-order terms can overfit noise and measurement error, leading to spurious spectral notches. In practice, therefore, the SH order must be limited based on the amount of sparse data available, and regularization or physics-based priors are added. However, SH interpolation will always struggle to capture higher frequency spectral content when the data is sparse.

### B. Learning-based Approaches

Recent advancements in ML have opened up promising avenues for HRTF personalization [51]–[53]. These data-driven methods aim to model the complex relationship between an individual's anatomical features and their corresponding HRTFs [54]–[57]. Methods based on autoencoder architectures [58]–[62] emphasize the frequency-domain characteristics of HRTFs by encoding them into compact latent representations. However, the upsampling performance has shown limited improvement over the algorithmic methods. Generative Adversarial Network (GAN) based models have demonstrated strong capabilities in reconstructing missing information from sparsely sampled HRTFs [35], [43]. By learning complex spatial and spectral patterns from a rich set of high-resolution HRTFs during training, these models can effectively infer plausible high-resolution outputs, even when the input measurements are limited. At least 4–5 measurements are required due to architectural constraints of the model, which limit its applicability in extremely sparse conditions. Moreover, LSD results often fail to align with perceptual evaluations for learning-based methods. In several cases, models that achieved strong performance in terms of LSD exhibited notably poor outcomes in perceptual assessments, highlighting a disconnect between objective metrics and subjective audio quality [63], [64].

### C. Transformer Models

Transformers introduced self-attention to model long-range dependencies without recurrence, instead using positional encodings to represent sequence order. This led to transformers rapidly becoming popular for natural language processing, along with becoming the state-of-the-art for many speech and audio applications [65]–[67]. The audio framework consists typically of a lightweight front end that converts waveforms or spectrograms into embeddings. Then the transformer is able to use these embeddings to capture the global context spatially, spectrally, and temporally. This flexibility motivates their use for HRTF upsampling, especially where preserving global spatial consistency is essential.

## III. METHOD

### A. Data Pre-processing

In general, HRTF data points are sampled on the surface of a sphere, making them inherently three-dimensional. To accommodate transformer architectures, which usually require one-dimensional sequential inputs, spherical harmonics transformation (SHT) [68] is utilized to project the HRTF data  $f(\theta, \phi)$  onto a series of orthogonal spherical harmonic basis functions and corresponding coefficients. The resulting SH coefficients  $F_l^m$  of degree  $l$  and order  $m$  are computed as:

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

where  $\theta$  and  $\phi$  represent the azimuth and elevation angles, respectively. In acoustic applications, the SH basis function is defined as:

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

where  $P_l^m(x)$  are the associated Legendre functions. The inverse SHT reconstructs the original HRTF function from its SH coefficients  $F_l^m$  via the following expression:

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

This SH representation offers multiple advantages: the coefficients naturally form a sequential structure compatible with transformer-based models. Each basis function represents a unique pattern of sound energy distribution in the space, and such a physically meaningful decomposition aligns well with the highly directional nature of HRTFs and facilitates more effective modeling and upsampling.

### B. Model Structure

As illustrated in Fig. 2, the proposed model adopts an encoder-decoder framework. The encoder takes low-resolution SH coefficients as input, extracts hierarchical features, and compresses them into a latent representation  $z$ . Subsequently, the decoder maps  $z$  back to high-resolution SH coefficients.

The structure of the encoder is depicted in Fig. 2(a). It consists of alternating transformer blocks and convolutional blocks, where transformer blocks capture global contextual relationships between coefficients via self-attention mechanisms, while convolutional blocks perform local feature extraction and feature map downsampling. Grouped query self-attention enables each head to focus on distinct spatial patterns, enhancing feature extraction in the transformer blocks.

Position encoding is vital in our task since low-degree spherical harmonics are typically more important than high-degree ones. However, since attention computation is permutation-invariant, explicit positional cues must be included. Therefore, rotary position embedding (RoPE) [69] is adopted. RoPE incorporates relative positional information by rotating query and key vectors, allowing the model to capture both local and global positional relationships. Compared to absolute position encodings [70], [71], RoPE better generalizes to variable input resolutions and provides a more natural way to model the ordering of SH coefficients. Consider a  $d$ -dimensional vector  $x$  at position index  $p$ , the rotary position embedding is applied by multiplying by a rotation matrix:

$$\text{RoPE}(x, p) = \begin{bmatrix} \cos\theta_{i,p} & -\sin\theta_{i,p} \\ \sin\theta_{i,p} & \cos\theta_{i,p} \end{bmatrix} \begin{bmatrix} x_{2i} \\ x_{2i+1} \end{bmatrix}, \quad \text{for } i = 0, 1, \dots, \frac{d}{2} - 1 \quad (4)$$

where  $\theta_{i,p}$  is the rotation angle.

$$Q' = \text{RoPE}(Q, p), \quad K' = \text{RoPE}(K, p), \quad (5)$$

$$\text{Attention}(Q', K', V) = \text{softmax}\left(\frac{Q' K'^\top}{\sqrt{d_k}}\right) V. \quad (6)$$

While layer normalization is commonly used in transformer architectures [70], [72], [73] for training stability, we empirically observe that the token scaling approach proposed in [74] demonstrates superior performance in our task. Unlike layer normalization, token scaling preserves the relative

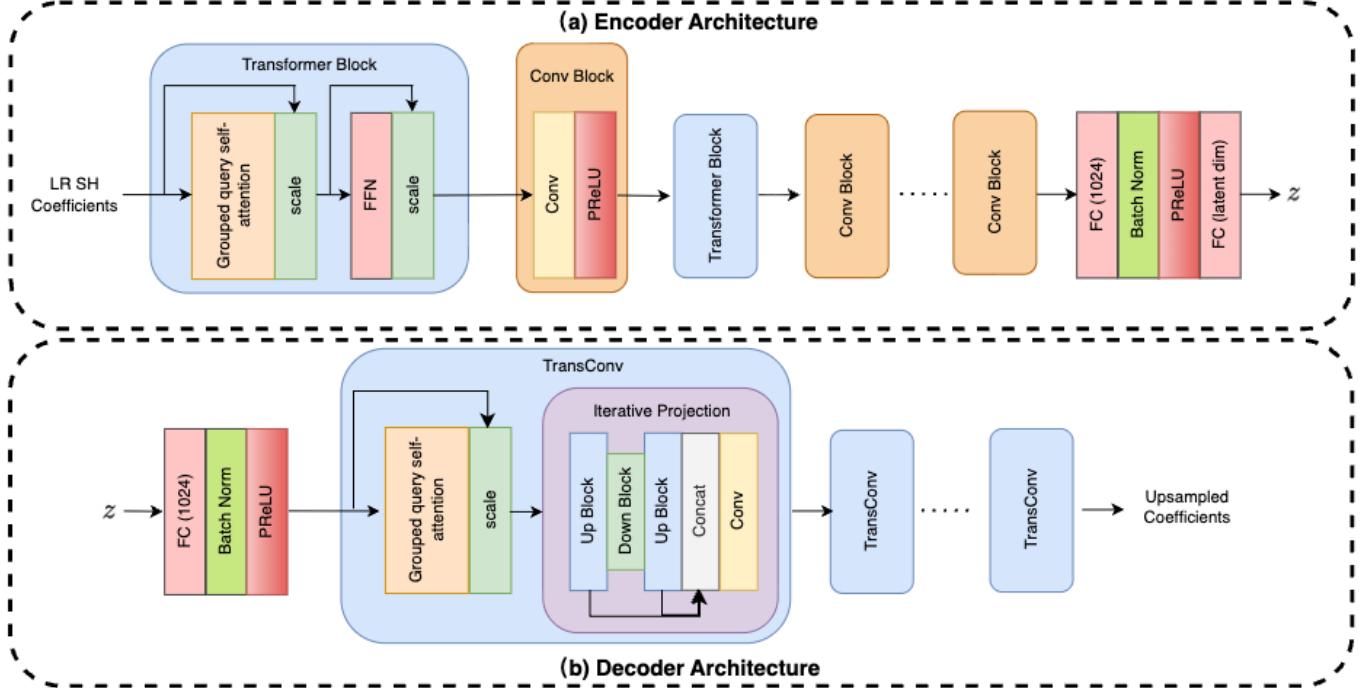


Fig. 2: The model architecture of our HRTFformer. The encoder integrates transformer layers with convolutional downsampling modules to progressively extract and compress spatial features from low-resolution SH coefficients into the latent representation. The decoder combines transformer layers with iterative projection units that perform upsampling.

energy distribution across frequency bins by avoiding mean subtraction, which is especially crucial in SH coefficient modeling because these coefficients are magnitude-sensitive and their relative scale carries physically and perceptually meaningful information. Token scaling can be written as:

$$y = \frac{x}{\sqrt{\frac{1}{C} \sum_{i=0}^{C-1} (x^i)^2 + \epsilon}}. \quad (7)$$

where  $x$  denotes the input feature,  $y$  represents the scaled output,  $C$  indicates the number of frequency channels, and  $\epsilon$  is a small constant for numerical stability.

The architecture of the decoder is illustrated in Fig. 2(b). The decoder, as shown in Fig. 2(b), maintains a similar structure with the encoder, but replaces convolutional blocks with iterative projection units for upsampling. This design adopts the iterative resolution refinement strategy proposed by [75], where repeated up- and downampling operations enable robust learning of the mapping between low- and high-resolution representations of SH coefficients. The arrows indicate residual connections. Transformer layers are also employed throughout the decoder to capture complex, long-range dependencies across different SH degrees and frequency components, which significantly improves interpolation and extrapolation performance, leading to finer high-resolution HRTF reconstruction results.

To address the upsampling challenge, we replace the standard feedforward layer in the transformer block with an iterative projection mechanism. This design adopts the iterative resolution refinement strategy proposed by [75], where repeated up- and downsampling operations enable robust learning of the mapping between low- and high-resolution representations of SH coefficients. The arrows indicate residual connections,

which further stabilize the training. Meanwhile, the attention module captures complex, long-range dependencies across different SH degrees and frequency components, which significantly improves interpolation and extrapolation performance, leading to finer high-resolution HRTF reconstruction results.

### C. Loss Functions

The overall loss function contains three terms: LSD, ILD, and neighbor dissimilarity loss (NDL). Therefore, the complete loss function can be shown as:

$$\mathcal{L} = \text{LSD} + \text{ILD} + \text{NDL}, \quad (8)$$

The LSD measures the discrepancy in magnitude between the reconstructed HRTF  $H_G$  and ground-truth high-resolution HRTF  $H_{HR}$ . It is defined as:

$$\text{LSD} = \frac{1}{N} \sum_{n=1}^N \sqrt{\frac{1}{W} \sum_{w=1}^W \left( 20 \log_{10} \frac{|H_{HR}(f_w, x_n)|}{|H_G(f_w, x_n)|} \right)^2}, \quad (9)$$

where  $N$  represents the total number of spatial positions, and  $W$  denotes the total number of frequency bins.  $H_{HR}$  and  $H_G$  are the targeted high-resolution HRTF and generated HRTF respectively.

The ILD quantifies the difference in magnitude between the left and right ear responses of an HRTF set. The ILD loss evaluates the deviation between the ILD values of the reconstructed and reference HRTFs, expressed as:

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

TABLE I: ITD, ILD, LSD evaluation results for sparsity level 3, 5, 19, and 100.

Method	Sparsity level 3			Sparsity level 5			Sparsity level 19			Sparsity level 100		
	ITD	ILD	LSD	ITD	ILD	LSD	ITD	ILD	LSD	ITD	ILD	LSD
GEP-GAN [43]	36.64	1.14	5.20	33.40	1.15	4.41	37.25	1.35	4.10	33.41	0.48	3.20
IOA3D [63]	22.84	1.00	4.67	16.00	0.75	4.90	13.95	0.69	3.21	<b>6.96</b>	0.41	<b>2.10</b>
SYT-FSP-AE [58]	24.66	1.28	4.42	18.32	1.07	4.36	21.38	0.91	3.25	17.83	0.76	2.21
Kalimotxo	30.46	0.92	4.49	31.39	0.72	4.85	25.08	0.81	3.29	21.18	0.73	3.06
AE-GAN	32.38	1.20	4.79	27.66	1.18	4.57	22.19	1.41	3.45	27.76	0.66	2.58
SH	78.61	6.05	9.96	77.14	5.44	10.35	62.75	1.68	5.43	47.67	0.44	3.38
Barycentric	49.05	7.50	8.56	47.86	4.54	8.33	45.37	1.76	4.79	41.32	0.55	3.20
HRTFformer(Ours)	<b>17.50</b>	<b>0.75</b>	<b>4.20</b>	<b>15.29</b>	<b>0.64</b>	<b>4.18</b>	<b>13.45</b>	<b>0.67</b>	<b>3.10</b>	19.36	<b>0.38</b>	3.15

The NDL is employed to encourage smooth change in the magnitude of the generated HRTF across the space. For each HRTF data point (treated as center), the neighbor dissimilarity quantifies the deviation between its value and the average of its four immediate neighbors (top, bottom, left, and right). Under the assumption of spatial continuity, this center-neighbor difference should be minimal. The neighbor dissimilarity loss quantifies the discrepancy between the neighbor dissimilarity patterns of the generated HRTF and the target HRTF, calculated as:

$$\mathcal{L}_{ND} = \frac{1}{N} \sum_{n=1}^N \left( \left( H_{HR}^{(n)} - \frac{1}{|\mathcal{K}(n)|} \sum_{k \in \mathcal{K}(n)} H_{HR}^{(k)} \right) - \left( H_G^{(n)} - \frac{1}{|\mathcal{K}(n)|} \sum_{k \in \mathcal{K}(n)} H_G^{(k)} \right) \right)^2. \quad (11)$$

where  $\mathcal{K}(n)$  represents the set of connected neighborhoods of position n,  $|\mathcal{K}(n)|$  is the number of neighboring points, which is 4 in our case. N is the total number of spatial positions.

## IV. EXPERIMENTS

### A. Implementation Details

In the experiments, the strides settings within the Conv Block were adjusted to accommodate varying input sizes according to different sparsity levels. The model was trained using a batch size of 8, and a learning rate of 0.0002 for 200 epochs and optimized using the Adam optimizer. All training was conducted on a single NVIDIA RTX 4090 GPU with 24GB memory.

### B. Tasks and Dataset

We evaluated our model on the SONICOM HRTF dataset [24], [76] under four sparsity levels—using 3, 5, 19, and 100 initial sampling points. To comprehensively assess performance, we employed both spatial cue metrics, including LSD, ILD, and ITD, which reflect the accuracy of binaural spatial cues crucial for externalization and spatial clarity, as well as perceptual localization metrics, which estimate how well listeners can localise sound sources based on the reconstructed HRTFs. Our model was compared against nine baselines, including algorithmic (non-ML) methods such as barycentric interpolation, SH and SUPDEq [77] methods, and deep learning models such as AE-GAN [60] and Kalimotxo [46].

### C. Result of Spatial Cue Evaluation

As shown in Table I, our HRTFformer consistently achieves the lowest errors across all three spatial cue metrics (ITD, ILD, and LSD) under the most challenging sparsity levels (3 and 5), significantly outperforming both learning-based and traditional interpolation baselines. This indicates that our model is robust in modeling spatial acoustic patterns even under extremely limited sampling conditions. Furthermore, at low sparsity levels (sparsity levels 100), HRTFformer obtains comparable results, where it performs on par with existing state-of-the-art methods. Although HRTFformer excels at high sparsity levels, its advantage diminishes with denser inputs, likely due to the transformer’s emphasis on global patterns over local details.

To better understand where these errors come from, Fig. 3 is a visualization of the spectral discrepancy across elevation and azimuth angles for four selected methods under four sparsity levels. Although all existing methods can capture HRTF spatial information and reconstruct it well at a low sparsity level. High sparsity levels represent more practical scenarios, as they significantly reduce the time and effort required from users during HRTF acquisition. As the sparsity level increases (moving leftward), baseline approaches show a noticeable rise in reconstruction errors, especially at sparsity level 3. In contrast, our model maintains significantly lower error levels, demonstrating its superior generalization in sparse settings due to its transformer-based architecture.

### D. Result of Perceptual Localisation Evaluation

Table II and III summarize the perceptual evaluation results. HRTFformer excels baselines by a large margin in both polar accuracy error as well as quadrant error on sparsity levels 3, 5, and 19, demonstrating its effectiveness in preserving high fidelity of reconstructed HRTFs, which is crucial for accurate sound localization. The comparable perceptual scores under sparsity level 100 suggest that spectral discrepancy does not always directly translate to perceptual differences. This observation is further supported by Fig. 5(a), where GEP-GAN, despite showing clear deviations in magnitude from the reference at sparsity level 3, still obtains competitive polar accuracy. At sparsity level 100, all methods manage to reconstruct spectral cues closely aligned with the reference for both ears, as shown in Fig. 5(b), reinforcing the notion that perceptual accuracy is not solely depended on spectral similarity.

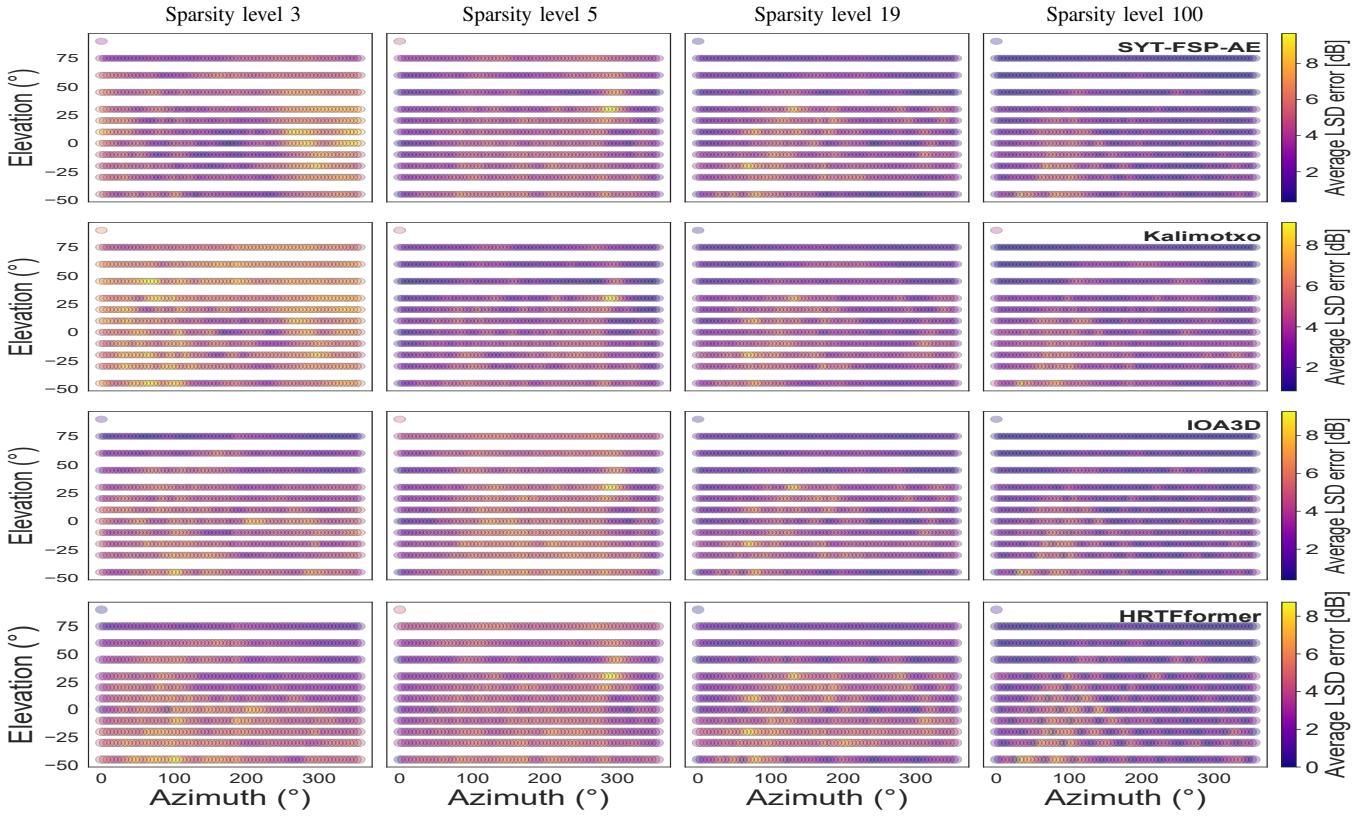


Fig. 3: LSD distributions for selected subject across various HRTF upsampling methods and sparsity levels. Top to bottom: SYT-FSP-AE, Kalimotxo, IOA3D, and HRTFFormer.

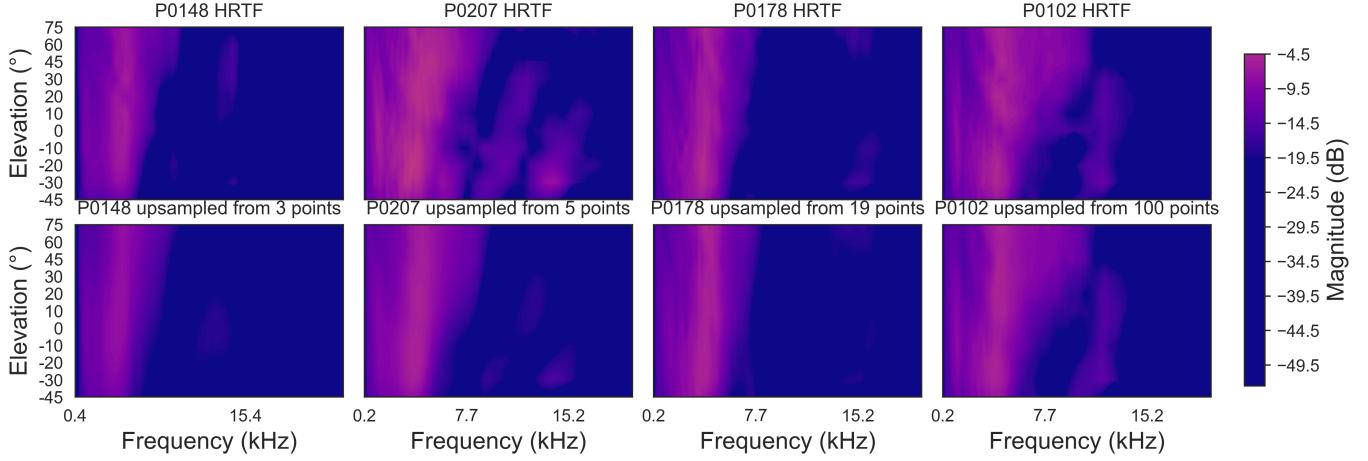


Fig. 4: Median plane spectra of example upsampled HRTFs using HRTFFormer compared to the original HRTFs for each upsampling factor.

In addition, Fig. 4 illustrates the spectral profiles of upsampled HRTFs generated by HRTFFormer, compared against the ground truth HRTFs for each sparsity level at azimuth of  $180^\circ$ .

#### E. Ablation Study

We conducted a series of ablation experiments to evaluate the impact of different encoder and decoder structures, position embedding methods, normalization techniques, and loss functions (Table IV).

**Model Structure.** As shown in Table IV, incorporating transformer modules in both encoder and decoder consistently

improves ITD, ILD, and LSD, showing the effectiveness of attention mechanisms in capturing dependencies of SH coefficients. In contrast, purely convolutional architectures primarily focus on local features, exhibit inferior performance.

**Position Embedding.** RoPE outperforms relative position bias [72], [74], [78] across all metrics except for ILD. This suggests that encoding positional information directly into the query and key vectors is more effective than adding a learnable bias term to the attention scores.

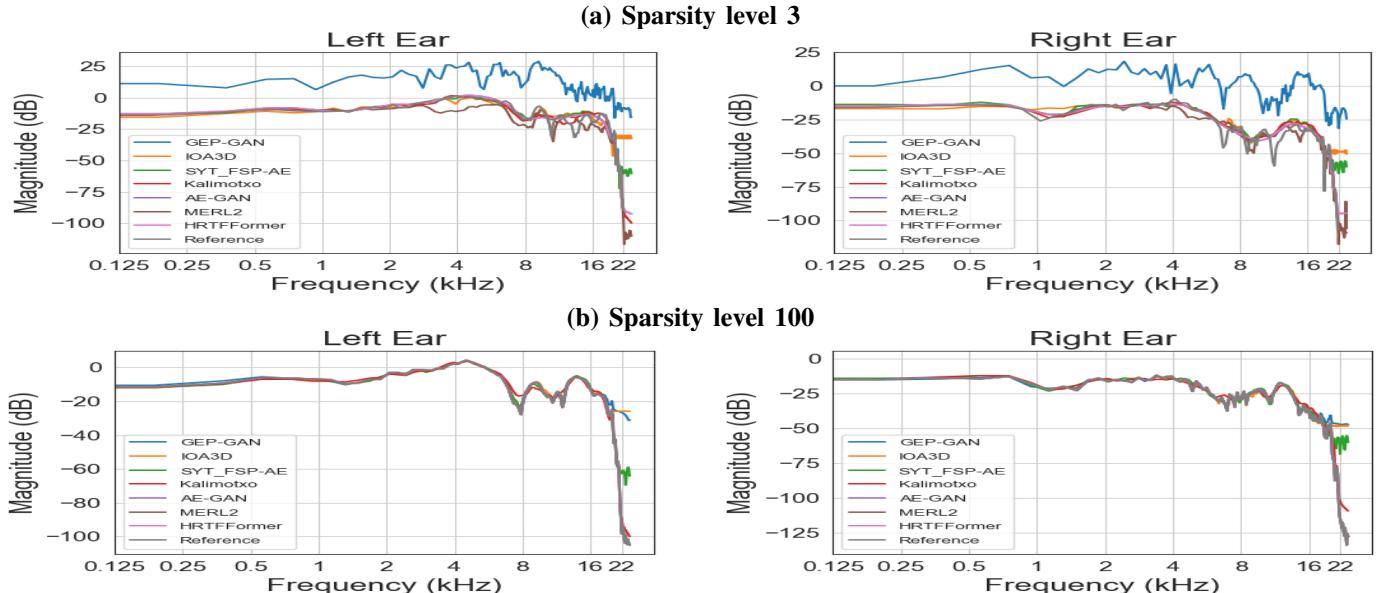
**Normalization.** Token scaling achieves the best results, yielding the lowest LSD and ITD along with a substantial

TABLE II: Perceptual evaluation results in sparsity level 3 and 5.

Method	Sparsity level 3				Sparsity level 5		
	Polar Accuracy Error	Polar RMS Error	Quadrant Error		Polar Accuracy Error	Polar RMS Error	Quadrant Error
GEP-GAN	5.66	47.60	16.35		5.22	46.06	28.17
IOA3D	13.75	41.79	18.20		14.74	41.01	23.35
Kalimotxo	10.45	41.72	18.08		17.50	41.42	25.44
AE-GAN	12.88	45.64	17.58		14.17	49.45	27.49
MERL2 [40]	11.53	38.55	13.50		7.77	35.53	16.39
SH	21.59	49.78	25.94		10.55	45.91	21.51
Barycentric	44.51	57.62	48.77		13.64	42.78	40.69
SUPDEq	16.53	47.15	27.96		21.15	38.36	20.22
HRTFformer(Ours)	<b>3.83</b>	<b>35.37</b>	<b>8.94</b>		<b>3.34</b>	<b>32.01</b>	<b>10.73</b>

TABLE III: Perceptual evaluation results in sparsity level 19 and 100.

Method	Sparsity level 19				Sparsity level 100		
	Polar Accuracy Error	Polar RMS Error	Quadrant Error		Polar Accuracy Error	Polar RMS Error	Quadrant Error
GEP-GAN	8.06	45.68	24.89		20.00	43.87	22.14
IOA3D	4.47	37.74	23.69		12.07	37.74	15.22
Kalimotxo	3.75	41.82	24.17		8.23	41.64	19.11
AE-GAN	7.41	39.96	25.72		15.44	42.36	16.98
MERL2	3.13	39.04	14.24		<b>4.27</b>	38.63	<b>12.56</b>
SH	6.95	45.11	29.59		10.16	39.63	20.05
Barycentric	9.77	41.68	31.48		15.23	43.75	23.86
SUPDEq	1.60	41.03	15.70		10.89	36.85	15.70
HRTFformer(Ours)	<b>0.19</b>	<b>30.26</b>	<b>10.22</b>		7.52	<b>35.78</b>	14.75

Fig. 5: Upsampled HRTFs for subject P0203 with two sparsity levels, and a source to the right ( $45^\circ$  azimuth,  $0^\circ$  elevation). The reference HRTF are shown for comparison.

improvement in polar accuracy, proving the previous discussion in Sec. III-B.

**Loss Functions.** Ablation results on loss components reveal that combining LSD and ILD captures spectral distribution characteristics better than MSE. Adding NDL further enables the model to account for local magnitude variations and prevent abrupt changes, resulting in more realistic reconstruction, as evidenced by the lowest errors in both spatial cue and most perceptual metrics.

## V. CONCLUSION AND FUTURE WORK

In this paper, we have proposed HRTFformer, a transformer-based model, to tackle the challenge of HRTF upsampling. It has been shown that by transforming HRTF data into the spherical harmonic domain and by leveraging attention mechanisms in transformer architecture, the model is able to learn the relationship between SH coefficients and, therefore, able to model the sound energy distribution pattern in the space effectively. A novel neighbor dissimilarity loss was also introduced to enforce

TABLE IV: Ablation study results in sparsity level 3

Component	Variations	Spatial Cue Evaluation			Perceptual Evaluation		
		ITD	ILD	LSD	Polar Accuracy Error	Polar RMS Error	Quadrant Error
Encoder	Resnet	18.27	0.77	4.32	5.54	43.05	<b>8.15</b>
	Transformer	<b>17.5</b>	<b>0.75</b>	<b>4.20</b>	<b>3.83</b>	<b>42.37</b>	8.94
Decoder	w/o Transformer	18.10	1.43	5.54	5.67	42.53	<b>6.81</b>
	w/ Transformer	<b>17.5</b>	<b>0.75</b>	<b>4.20</b>	<b>3.83</b>	<b>42.37</b>	8.94
Position Embedding	Relative Position Bias	18.25	<b>0.66</b>	4.23	6.08	42.82	8.94
	ROPE	<b>17.5</b>	0.75	<b>4.20</b>	<b>3.83</b>	<b>42.37</b>	<b>8.39</b>
Normalization	LayerNorm	17.74	<b>0.66</b>	4.25	5.65	43.05	9.81
	BatchNorm	18.83	0.76	4.41	7.17	42.60	<b>7.17</b>
	Token Scaling	<b>17.5</b>	0.75	<b>4.20</b>	<b>3.83</b>	<b>42.37</b>	8.94
Loss Functions	MSE	18.95	0.89	4.87	8.14	42.46	<b>7.13</b>
	LSD+ILD	17.98	0.91	4.26	5.21	42.85	9.73
	LSD+ILD+NDL	<b>17.5</b>	<b>0.75</b>	<b>4.20</b>	<b>3.83</b>	<b>42.37</b>	8.94

spatial continuity in the HRTF magnitude spectrum across adjacent positions to achieve a more realistic HRTF reconstruction. The statistical results have suggested that HRTFformer is not only able to outperform other state-of-the-art methods in terms of objective metrics (LSD, ILD, and ITD) but also excels in a perceptual evaluation by a large margin, highlighting the model's effectiveness for use in real-life applications.

In future work, we plan to confirm our perceptual results with subjective evaluations using real human listeners to assess spatial realism, externalization, and personalized localization. This will overcome a current limitation of our results, which is that they rely on a perceptual model. It should also be noted that due to the nature of HRTF measurements being costly and time-consuming, datasets are limited, and, therefore, data diversity could have the potential to affect generalization to unseen subjects. To overcome this, we propose a future work that uses synthetic HRTFs for use in training via means of transfer learning, as this will likely help improve model generalization to unseen subjects/conditions.

## REFERENCES

- J. Herre and S. Disch, "Mpeg-i immersive audio-reference model for the virtual/augmented reality audio standard," *Journal of the Audio Engineering Society*, vol. 71, no. 5, pp. 229–240, 2023.
- S. R. Quackenbush and J. Herre, "Mpeg standards for compressed representation of immersive audio," *Proceedings of the IEEE*, vol. 109, no. 9, pp. 1578–1589, 2021.
- M. A. Steadman, C. Kim, J.-H. Lestang, D. F. Goodman, and L. Picinali, "Short-term effects of sound localization training in virtual reality," *Scientific Reports*, vol. 9, no. 1, p. 18284, 2019.
- X. Sun, "Immersive audio, capture, transport, and rendering: A review," *APSIPA Transactions on Signal and Information Processing*, vol. 10, p. e13, 2021.
- D. Brahimaj, E. Vezzoli, F. Giraud, and B. Semail, "Enhancing object localization in vr: Tactile-based hrtf and vibration headphones for spatial haptic feedback," *IEEE Transactions on Haptics (ToH)*, 2024.
- P. M. Silva, T. N. Pappas, J. Atkins, and J. E. West, "Perceiving graphical and pictorial information via hearing and touch," *IEEE Transactions on Multimedia*, vol. 18, no. 12, pp. 2432–2445, 2016.
- L. Picinali and B. F. Katz, "System-to-user and user-to-system adaptations in binaural audio," in *Sonic interactions in virtual environments*. Springer International Publishing Cham, 2022, pp. 115–143.
- B. C. Moore, *An introduction to the psychology of hearing*. Brill, 2012.
- V. Bruschi, L. Grossi, N. A. Dourou, A. Quattrini, A. Vancheri, T. Leidi, and S. Cecchi, "A review on head-related transfer function generation for spatial audio," *Applied Sciences*, vol. 14, no. 23, p. 11242, 2024.
- A. Mouchtaris, P. Reveliotis, and C. Kyriakakis, "Inverse filter design for immersive audio rendering over loudspeakers," *IEEE transactions on multimedia*, vol. 2, no. 2, pp. 77–87, 2002.
- C. T. Jin, P. Guillon, N. Epain, R. Zolfaghari, A. Van Schaik, A. I. Tew, C. Hetherington, and J. Thorpe, "Creating the sydney york morphological and acoustic recordings of ears database," *IEEE Transactions on Multimedia*, vol. 16, no. 1, pp. 37–46, 2013.
- P. Stitt, L. Picinali, and B. F. Katz, "Auditory accommodation to poorly matched non-individual spectral localization cues through active learning," *Scientific reports*, vol. 9, no. 1, p. 1063, 2019.
- X. Lu, Y. Wang, J. Sang, and C. Zheng, "Bicg: Binaural cue generation from unified hrtf datasets," in *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2025, pp. 1–5.
- S. Werner, F. Klein, T. Mayenfels, and K. Brandenburg, "A summary on acoustic room divergence and its effect on externalization of auditory events," in *2016 Eighth International Conference on Quality of Multimedia Experience (QoMEX)*. IEEE, 2016, pp. 1–6.
- I. Engel, D. L. Alon, P. W. Robinson, and R. Mehra, "The effect of generic headphone compensation on binaural renderings," in *Audio Engineering Society Conference: 2019 AES International Conference on Immersive and Interactive Audio*. Audio Engineering Society, 2019.
- J. Zhao, D. Yao, and J. Li, "Cross-dataset head-related transfer function harmonization based on perceptually relevant loss function," *IEEE Open Journal of Signal Processing*, 2025.
- M. Cuevas-Rodríguez, L. Picinali, D. González-Toledo, C. Garre, E. de la Rubia-Cuestas, L. Molina-Tanco, and A. Reyes-Lecuona, "3d tune-in toolkit: An open-source library for real-time binaural spatialisation," *PloS one*, vol. 14, no. 3, p. e0211899, 2019.
- D. Yao, J. Zhao, L. Cheng, J. Li, X. Li, X. Guo, and Y. Yan, "An individualization approach for head-related transfer function in arbitrary directions based on deep learning," *JASA Express Letters*, vol. 2, no. 6, 2022.
- P. Coleman, A. Franck, J. Francombe, Q. Liu, T. De Campos, R. J. Hughes, D. Menzies, M. F. S. Gálvez, Y. Tang, J. Woodcock *et al.*, "An audio-visual system for object-based audio: from recording to listening," *IEEE Transactions on Multimedia*, vol. 20, no. 8, pp. 1919–1931, 2018.
- A. O. T. Hogg, V. W. Neo, S. Weiss, C. Evers, and P. A. Naylor, "A polynomial eigenvalue decomposition music approach for broadband sound source localization," in *2021 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2021, pp. 326–330.
- C. Guezenoc and R. Seguier, "Hrtf individualization: A survey," *arXiv preprint arXiv:2003.06183*, 2020.
- M. Zhang, X. Wu, and T. Qu, "Individual distance-dependent hrtfs modeling through a few anthropometric measurements," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 401–405.
- N. Marggraf-Turley, N. H. Pontoppidan, and L. Picinali, "The impact of individual head-related transfer function augmentation on spatial release from masking," *Hearing Research*, p. 109402, 2025.
- I. Engel, R. Daugintis, T. Vicente, A. O. T. Hogg, J. Pauwels, A. J. Tournier, and L. Picinali, "The sonicom hrtf dataset," *journal of the audio engineering society*, vol. 71, no. 5, pp. 241–253, 2023.

- [25] M. Dinakaran, F. Brinkmann, S. Harder, R. Pelzer, P. Grosche, R. R. Paulsen, and S. Weinzierl, "Perceptually motivated analysis of numerically simulated head-related transfer functions generated by various 3d surface scanning systems," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 551–555.
- [26] M. Zhang, Z. Ge, T. Liu, X. Wu, and T. Qu, "Modeling of individual hrtfs based on spatial principal component analysis," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 785–797, 2020.
- [27] F. Grijalva, L. Martini, D. Florencio, and S. Goldenstein, "A manifold learning approach for personalizing hrtfs from anthropometric features," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 3, pp. 559–570, 2016.
- [28] Y. Wang, Y. Zhang, Z. Duan, and M. Bocko, "Global hrtf personalization using anthropometric measures," in *Audio Engineering Society Convention: 2020 AES International Conference on Audio for Virtual and Augmented Reality*. Audio Engineering Society, 2020.
- [29] Y. Zhang, Y. Wang, M. Bocko, and Z. Duan, "Grid-agnostic personalized head-related transfer function modeling with neural fields," *The Journal of the Acoustical Society of America*, vol. 153, no. 3\_supplement, pp. A125–A125, 2023.
- [30] N. Marggraf-Turley, M. Lovedee-Turner, and E. De Sena, "Hrtf recommendation based on the predicted binaural colouration model," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 1106–1110.
- [31] H. S. Braren and J. Fels, "Towards child-appropriate virtual acoustic environments: A database of high-resolution hrtf measurements and 3d-scans of children," *International Journal of Environmental Research and Public Health*, vol. 19, no. 1, p. 324, 2021.
- [32] G.-T. Lee, S.-M. Choi, B.-Y. Ko, and Y.-H. Park, "Hrtf measurement for accurate sound localization cues," *arXiv preprint arXiv:2203.03166*, 2022.
- [33] K. Niwa, Y. Hioka, and H. Uematsu, "Efficient audio rendering using angular region-wise source enhancement for 360 video," *IEEE Transactions on Multimedia*, vol. 20, no. 11, pp. 2871–2881, 2018.
- [34] P. Siripornpitak, I. Engel, I. Squires, S. J. Cooper, and L. Picinali, "Spatial up-sampling of hrtf sets using generative adversarial networks: A pilot study," *Frontiers in Signal Processing*, vol. 2, p. 904398, 2022.
- [35] J. Zhao, D. Yao, and J. Li, "Head-related transfer function upsampling with spatial extrapolation features," *IEEE Transactions on Audio, Speech and Language Processing*, 2025.
- [36] D. Bau, J. M. Arend, and C. Pörschmann, "Estimation of the optimal spherical harmonics order for the interpolation of head-related transfer functions sampled on sparse irregular grids," *Frontiers in Signal Processing*, vol. 2, p. 884541, 2022.
- [37] M. Zhang, T. Guan, L. Chen, T. Fu, D. Su, and T. Qu, "Individualized hrtf-based binaural renderer for higher-order ambisonics," in *Audio Engineering Society Convention 150*. Audio Engineering Society, 2021.
- [38] R. Fernandez Martinez, P. Jimbert, E. M. Sumner, M. Riedel, and R. Unnthorsson, "Prediction of head related transfer functions using machine learning approaches," vol. 5, no. 1, pp. 254–267, 2023.
- [39] R. A. McCarthy, Y. Zhang, S. A. Verburg, W. F. Jenkins, and P. Gerstoft, "Machine learning in acoustics: A review and open-source repository," *arXiv preprint arXiv:2507.04419*, 2025.
- [40] Y. Masuyama, G. Wichern, F. G. Germain, C. Ick, and J. Le Roux, "Retrieval-augmented neural field for hrtf upsampling and personalization," in *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2025, pp. 1–5.
- [41] B.-Y. Ko, G.-T. Lee, H. Nam, and Y.-H. Park, "Prtnet: Hrtf individualization for accurate spectral cues using a compact prtf," *IEEE Access*, vol. 11, pp. 96119–96130, 2023.
- [42] X. Lu, Y. Chen, Z. Chen, J. Wang, M. Liu, H. Hu, C. Zheng, S. Bleeck, and J. Sang, "Deep learning for personalized binaural audio reproduction," *arXiv preprint arXiv:2509.00400*, 2025.
- [43] A. O. T. Hogg, M. Jenkins, H. Liu, I. Squires, S. J. Cooper, and L. Picinali, "Hrtf upsampling with a generative adversarial network using a gnomonic equiangular projection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024.
- [44] J. Xia and W. Zhang, "Upmix b-format ambisonic room impulse responses using a generative model," *Applied Sciences*, vol. 13, no. 21, p. 11810, 2023.
- [45] R. Niu, S. Koyama, and T. Nakamura, "Head-related transfer function individualization using anthropometric features and spatially independent latent representation," *arXiv preprint arXiv:2508.16176*, 2025.
- [46] A. O. Hogg, R. Barumerli, R. Daugintis, K. C. Poole, F. Brinkmann, L. Picinali, and M. Geronazzo, "Listener acoustic personalisation challenge-lap24: Head-related transfer function upsampling," *IEEE Open Journal of Signal Processing*, 2025.
- [47] J. Li, X. Su, and K. Zhao, "Barycentric interpolation collocation algorithm to solve fractional differential equations," *Mathematics and Computers in Simulation*, vol. 205, pp. 340–367, 2023.
- [48] J.-P. Berrut and L. N. Trefethen, "Barycentric lagrange interpolation," *SIAM review*, vol. 46, no. 3, pp. 501–517, 2004.
- [49] R. Suda and M. Takami, "A fast spherical harmonics transform algorithm," *Mathematics of computation*, vol. 71, no. 238, pp. 703–715, 2002.
- [50] X. Hu, J. Li, L. Picinali, and A. O. T. Hogg, "A machine learning approach for denoising and upsampling hrtfs," *Proceedings of the 33rd European Signal Processing Conference*, 2025.
- [51] X. Chen, F. Ma, Y. Zhang, A. Bastine, and P. N. Samarasinghe, "Head-related transfer function interpolation with a spherical cnn," *arXiv preprint arXiv:2309.08290*, 2023.
- [52] Y. Zhang, Y. Wang, and Z. Duan, "Hrtf field: Unifying measured hrtf magnitude representation with neural fields," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.
- [53] Y. Ito, T. Nakamura, S. Koyama, S. Sakamoto, and H. Saruwatari, "Spatial upsampling of head-related transfer function using neural network conditioned on source position and frequency," *IEEE Open Journal of Signal Processing*, 2025.
- [54] Y. Masuyama, G. Wichern, F. G. Germain, Z. Pan, S. Khurana, C. Hori, and J. Le Roux, "Niirf: Neural iir filter field for hrtf upsampling and personalization," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 1016–1020.
- [55] I. D. Gebru, D. Marković, A. Richard, S. Krenn, G. A. Butler, F. De la Torre, and Y. Sheikh, "Implicit hrtf modeling using temporal convolutional networks," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 3385–3389.
- [56] J. W. Lee, S. Lee, and K. Lee, "Global hrtf interpolation via learned affine transformation of hyper-conditioned features," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.
- [57] F. Ma, T. D. Abhayapala, P. N. Samarasinghe, and X. Chen, "Spatial upsampling of head-related transfer functions using a physics-informed neural network," *arXiv preprint arXiv:2307.14650*, 2023.
- [58] Y. Ito, T. Nakamura, S. Koyama, and H. Saruwatari, "Head-related transfer function interpolation from spatially sparse measurements using autoencoder with source position conditioning," in *2022 International Workshop on Acoustic Signal Enhancement (IWAENC)*. IEEE, 2022, pp. 1–5.
- [59] 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," *The Journal of the Acoustical Society of America*, vol. 153, no. 1, pp. 248–259, 2023.
- [60] X. Hu, J. Li, L. Picinali, and A. O. T. Hogg, "HRTF spatial upsampling in the spherical harmonics domain employing a generative adversarial network," *Proceedings of the International Conference on Digital Audio Effects*, 2024.
- [61] W. Chen and X. Wei, "Spatial interpolation of head-related transfer functions using a physics-informed autoencoder," *Multimedia Systems*, vol. 31, no. 3, p. 247, 2025.
- [62] X. Hu, J. Li, L. Picinali, and A. Hogg, "Head-related transfer function upsampling using an autoencoder-based generative adversarial network with evaluation framework," *Journal of the Audio Engineering Society*, 2025.
- [63] M. Geronazzo, L. Picinali, A. Hogg, R. Barumerli, K. Poole, R. Daugintis, J. Pauwels, S. Ntalampiras, G. McLachlan, and F. Brinkmann, "Technical report: Sonicom/ieee listener acoustic personalisation (lap) challenge-2024," 2024.
- [64] D. Fantini, M. Geronazzo, F. Avanzini, and S. Ntalampiras, "A survey on machine learning techniques for head-related transfer function individualization," *IEEE Open Journal of Signal Processing*, 2025.
- [65] K. Zaman, K. Li, M. Sah, C. Direkoglu, S. Okada, and M. Unoki, "Transformers and audio detection tasks: An overview," *Digital Signal Processing*, vol. 158, p. 104956, 2025.
- [66] L. Dong, S. Xu, and B. Xu, "Speech-transformer: a no-recurrence sequence-to-sequence model for speech recognition," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018, pp. 5884–5888.
- [67] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu *et al.*, "Conformer: Convolution-augmented transformer for speech recognition," *arXiv preprint arXiv:2005.08100*, 2020.
- [68] M. J. Mohlenkamp, "A fast transform for spherical harmonics," *Journal of Fourier analysis and applications*, vol. 5, pp. 159–184, 1999.

- [69] J. Su, M. Ahmed, Y. Lu, S. Pan, W. Bo, and Y. Liu, "Roformer: Enhanced transformer with rotary position embedding," *Neurocomputing*, vol. 568, p. 127063, 2024.
- [70] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.
- [71] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.
- [72] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022.
- [73] P. Zhang, X. Dai, J. Yang, B. Xiao, L. Yuan, L. Zhang, and J. Gao, "Multi-scale vision longformer: A new vision transformer for high-resolution image encoding," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 2998–3008.
- [74] Y. Jiang, S. Chang, and Z. Wang, "Transgan: Two pure transformers can make one strong gan, and that can scale up," *Advances in Neural Information Processing Systems*, vol. 34, pp. 14 745–14 758, 2021.
- [75] M. Haris, G. Shakhnarovich, and N. Ukita, "Deep back-projection networks for super-resolution," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 1664–1673.
- [76] K. C. Poole, J. Meyer, V. Martin, R. Daugintis, N. Marggraf-Turley, J. Webb, L. Pirard, N. La Magna, O. Turvey, and L. Picinali, "The extended sonicom hrtf dataset and spatial audio metrics toolbox," *arXiv preprint arXiv:2507.05053*, 2025.
- [77] J. M. Arend and C. Pörschmann, *Spatial upsampling of sparse head-related transfer function sets by directional equalization-influence of the spherical sampling scheme*. Universitätsbibliothek der RWTH Aachen Aachen, Germany, 2019.
- [78] P. Shaw, J. Uszkoreit, and A. Vaswani, "Self-attention with relative position representations," *arXiv preprint arXiv:1803.02155*, 2018.



**Xuyi Hu** (Student Member, IEEE) is a PhD student at University of Cambridge. 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. More information about current research projects can be found at: <https://georgehu.com/>.



**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.



**Shaojie Zhang** (Member, IEEE) received the B.Eng. degree in electronics and electrical engineering from University College London in 2023, and the M.Phil. degree from the Department of Engineering, University of Cambridge in 2024, where he is currently pursuing the Ph.D. degree. His research interests include the applications of molecular communication systems and integrated sensing and communication within molecular communication.



**Stefan Goetz** (Member, IEEE) studied physics and engineering. He received his undergraduate and graduate degrees from Technische Universität München (TU München), Munich, Germany, where he also conducted Ph.D. research on medical power electronics, partly at Columbia University, New York, NY, USA. He has worked in the automotive industry on electric drivetrains, machines, power electronics and vehicle architecture. He also contributed to the development of automotive chargers and grid integration technologies deployed worldwide. His research focuses on high-power electronics and magnetics for drive and medical applications, and on integrative power electronics for microgrids and e-vehicles.



**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/>.



**Ozgur B. Akan** (Fellow, IEEE) received the Ph.D. degree from the School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, in 2004. He is currently the Head of the Internet of Everything Group, Department of Engineering, University of Cambridge, U.K., and the Director of the Centre for Next-Generation Communications, Koç University, Türkiye. His research interests include wireless, nano, and molecular communications and Internet of Everything.



**Aidan O. T. Hogg** 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/>