



D-GAT: Dual Graph Attention Network for Global HRTF Interpolation

Junsheng Hu¹, Shaojie Li¹, Qintuya Si², De Hu^{1,*}

¹College of Computer Science, Inner Mongolia University, China

²College of Electronic and Information Engineering, Inner Mongolia University, China

hujunsheng@mail.imu.edu.cn, lishaojie@mail.imu.edu.cn, sigty@imu.edu.cn,
cshood@imu.edu.cn

Abstract

To achieve 3D audio rendering, high-quality Head-Related Transfer Functions (HRTFs) are essential. As measuring HRTFs is time-consuming and tedious, spatial interpolation is often adopted to generate high-resolution HRTFs from low-resolution ones. In this paper, we propose a Dual-Graph Attention Network (D-GAT) for HRTF upsampling. Specifically, we first design a branch of GAT to learn the relationship among HRTFs from adjacent points. In addition, we introduce another branch of GAT to find a mapping from physical features (including the absolute target position and the anthropometric characteristics) to HRTFs. By combining such two GATs in a parallel architecture, the D-GAT is built. Furthermore, a dynamic edge weighting mechanism is adopted in the D-GAT, which allows the model to learn geometry relationships among vertices more flexibly. Experimental results demonstrate the efficacy of the proposed D-GAT in accurately predicting HRTFs, yielding state-of-the-art performance.

Index Terms: Head-related transfer function, interpolation, graph attention networks.

1. Introduction

There is a growing demand for 3D audio rendering in our daily life, such as teleconference systems [1], virtual reality (VR) [2, 3] or augmented reality (AR) [4, 5], and hearing assistive devices [6, 7], to name but a few. For headphone-based 3D audio rendering, the key component is head-related transfer functions (HRTFs), which contain binaural spatial cues for simulating the effects of sound fields on two ears [8]. As measuring HRTFs is time-consuming and requires an expensive acoustic lab setting, the size of existing HRTF datasets is usually limited. To that end, several spatial interpolation approaches have been developed over the past decades, aiming to generate high-quality high-resolution HRTFs from a small dataset [9, 10].

In the early stages, traditional interpolation methods have been widely applied to predict HRTF at an arbitrary position by interpolating HRTFs with known distributions [11]. Two common methods are barycentric interpolation [12, 13, 14] and spherical harmonic interpolation [15, 16], by weighting and summing the HRTFs from a few nearest neighbors around the target point. These methods achieve good performance when interpolating dense HRTF measurements, since the HRTF of the target position is strongly correlated with those HRTFs from its nearest neighbors. For a sparse HRTF dataset (the scale of existing databases in some acoustic environments is very limited), however, their interpolation accuracy decreases signifi-

cantly. This is because the similarity of HRTFs between the target point and its neighbors decreases with increasing distances between adjacent points (Figure 1).

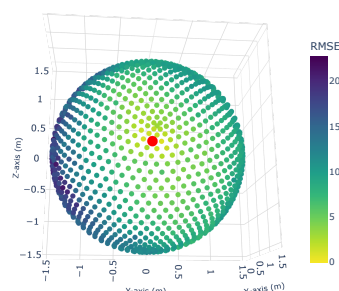


Figure 1: A randomly selected target point (red point) from the HUTUBS database, and its surrounding points with an angular distance less than 0.5π . The color bar indicates the RMSE of HRTF (from left ear) between the target point and other points.

To further improve the accuracy of HRTF interpolation, data-driven methods have gained significant attention in recent years [17, 18]. According to HRTF generation models, recent advances can be divided into two main classes, i.e., direct mapping (DM) [19, 20] and indirect mapping (IM) [21, 22]. The former aims to directly find a mapping from physical features (including the absolute target position and the anthropometric characteristics) to HRTFs, while the latter generates HRTFs via spatial interpolation by learning high-accuracy interpolation rules. In general, DM involves the challenge of mapping low-dimensional physical features to high-dimensional HRTFs, leading to limited accuracy. To address this, HRTFs were predicted in a frequency-independent manner [23, 24], which can learn a better relationship between physical characteristics and HRTFs. In contrast, IM generates HRTFs by mapping those HRTFs from neighboring points to target points, thus it maintains a higher precision. However, the size of neighboring sets around the target point is often unfixed, especially for an uneven database, which conflicts with the requirement of fixed input sizes in most deep learning models. To this end, the state of the art [25] utilized a repeated sampling strategy to select a fixed number of neighbors around the target point. Still, some information-rich points may be missed in such a repeated sampling procedure, resulting in lower stability.

In this work, we propose a Dual-Graph Attention Network (D-GAT) for global HRTF interpolation, which integrates the DM and the IM in a parallel architecture, delivering the advantages of both DM and IM approaches. Specifically, a branch of

* Corresponding author. The source code is available at <https://doi.org/10.5281/zenodo.15522810>.

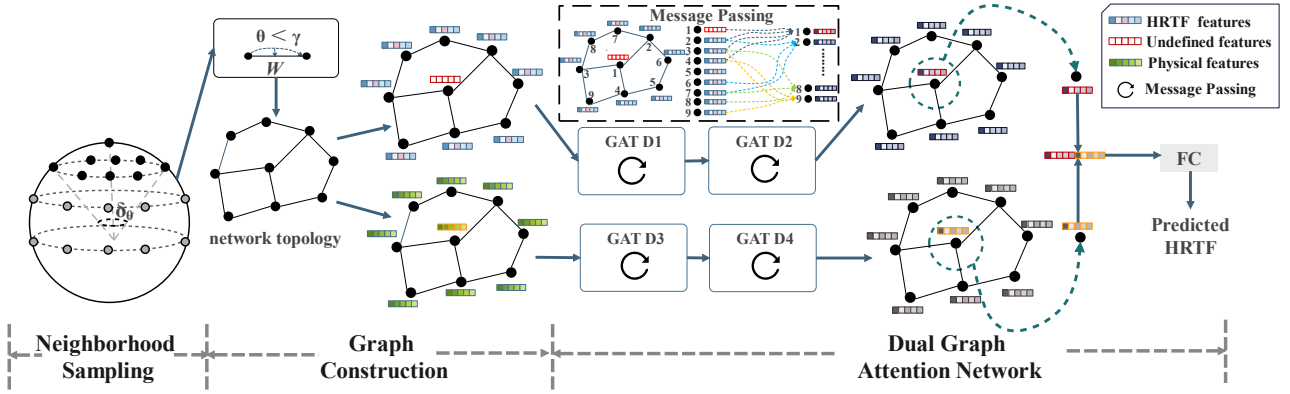


Figure 2: The proposed network architecture, where “FC” refers to the fully connected layer.

GAT is first constructed for the IM, which learns the mapping of HRTFs among adjacent points. Meanwhile, another branch of GAT is designed for the DM, which directly maps the physical features into HRTFs. By combining these two GATs in a linear fashion, their outputs can be fused efficiently in the D-GAT. To the best of our knowledge, this is the first attempt to combine the HRTF interpolation task with graph neural networks (GNNs). Compared with the existing DM and IM approaches, D-GAT supports unfixed input sizes and achieves higher accuracy and stability. Numerical experiments validate its efficacy.

2. PROBLEM STATEMENT

Our goal is to generate HRTFs of arbitrary positions using a neural network. As the phase information of HRTFs can be effectively reconstructed using the minimum-phase approximation [14, 17, 24], we focus only on the magnitude of HRTFs. Let $\mathcal{H}_p^{(a)} \in \mathbb{R}^K$ be the HRTF magnitude of the target position $p \in \mathbb{R}^3$ with K denoting the number of frequency bins, and $a \in \mathbb{R}^J$ denoting the anthropometric characteristic with the dimension of J . Then, an IM Φ between $\mathcal{H}_p^{(a)}$ and its neighboring measurements can be constructed as [25]

$$\Phi : \left(\mathcal{H}_{q \in \mathcal{N}_p}^{(a)}, \mathcal{N}_p, p, a \right) \mapsto \mathcal{H}_p^{(a)} \in \mathbb{R}^K, \quad (1)$$

where \mathcal{N}_p is the set of neighbors corresponding to the target point p , which is defined by

$$\mathcal{N}_p = \{0 < d_E(q, p) < \delta_d\}, \quad (2)$$

where $d_E(q, p)$ represents the Euclidean distance between points q and p , and $\delta_d > 0$ is a threshold. As shown in Figure 1, $\mathcal{H}_p^{(a)}$ is only significantly correlated with a few HRTFs from the nearest neighbors of p . That is, δ_d can be set to a small constant to reduce computational complexity.

In general, determining Φ with a neural network requires that the size of inputs, i.e., $|\mathcal{N}_p|$, is fixed. However, $|\mathcal{N}_p|$ in Equation (2) varies with p in an unevenly distributed dataset, resulting in a non-fixed size for inputs. To this end, [25] employs the repeated sampling strategy to select a fixed number of neighbors from \mathcal{N}_p . However, this brings two additional issues as follows: *on the one hand, some points in $|\mathcal{N}_p|$ may be sampled repeatedly, causing data redundancy; on the other hand, some important points may be missed in repeated sampling, resulting in information loss.*

3. PROPOSED METHOD

In this work, the GNN is adopted for HRTF upsampling, due to the fact that

- GNN supports inputs with non-fixed sizes and thus can avoid the issues in the end of Sec. 2;
- GNN is well-suited for representing spatial relationship among node with local similarities (e.g., Figure 1).

The proposed network architecture is depicted in Figure 2. We first adopt the target point and its neighbors to establish a graph, where the edge set is determined via angle distance. Then, such a graph is split into two separate graphs, and their node representations are the HRTF feature and the physical feature, respectively. Next, we use two branches of graph Attention networks (GATs) [26] as our backbone modules to achieve IM and DM mappings, where each branch concatenates two GAT layers. Finally, the outputs of two branches are fused by a fully connected layer.

3.1. Neighborhood Sampling

The HRTFs in existing datasets [27, 28] are often measured on a spherical surface with a constant radius centered on the human head. For this reason, instead of using the Euclidean distance as used in Equation (2), the angular distance is employed here to determine the neighboring sets, i.e.,

$$\mathcal{N}_p = \{0 < \theta(q, p) < \delta_\theta\}, \quad (3)$$

where $\theta(q, p) = \arccos\left(\frac{q^\top p}{\|q\| \|p\|}\right)$ represents the angular distance and δ_θ the threshold, $\|\cdot\|$ is the ℓ_2 norm, and the superscript \top denotes the transpose operation.

3.2. Graph Construction

Based on Sec. 2, we define a new neighboring set for the target point p , which consists of p and \mathcal{N}_p , i.e., $\mathcal{N}_p^+ = \mathcal{N}_p \cup \{p\}$. Then, the vertex set $\mathcal{V}_p^{(a)}$ of \mathcal{N}_p^+ is expressed as

$$\mathcal{V}_p^{(a)} = \{\mathcal{H}_q^{(a)} \mid q \in \mathcal{N}_p\} \cup \mathcal{H}_p^{(a)}, \quad (4)$$

where the unknown $\mathcal{H}_p^{(a)}$ is initialized by an all-ones vector of size K due to the lack of any prior information. It can also be initialized with traditional interpolation methods if possible. In addition, the edge set of the graph can be constructed as

$$\mathcal{E}_p^{(a)} = \{(q, b) \mid 0 < \theta(q, b) < \gamma, q, b \in \mathcal{N}_p^+\}, \quad (5)$$

where γ is a threshold value that is smaller than δ_θ in (3). This is because a larger δ_θ allows \mathcal{N}_p^+ to include more points that potentially affect $\mathcal{H}_p^{(a)}$, while a smaller γ allows the network to study the relationship of graph features among closer nodes.

From Figure 1, we infer that the smaller the distance among the nodes, the greater the similarity among their HRTFs. To this end, we introduce a distance-dependent Gaussian kernel function (GKF) for weighting edges, which emphasizes the influence among neighboring nodes. To be specific, the set of edge weights $\mathcal{W}_p^{(a)}$ is defined as

$$\mathcal{W}_p^{(a)} = \left\{ \exp\left(-\frac{d_E(q,b)^2}{2\sigma^2}\right) \mid (q,b) \in \mathcal{E}_p^{(a)} \right\}, \quad (6)$$

where σ is the bandwidth parameter of the GKF. Based on the above, we construct a graph $\mathcal{G}_p^{(a)}$ for the target point p as

$$\mathcal{G}_p^{(a)} = \{\mathcal{V}_p^{(a)}, \mathcal{E}_p^{(a)}, \mathcal{W}_p^{(a)}\}. \quad (7)$$

Note that such a graph relies on the anthropometric characteristics a , indicating that the graph in (7) may vary with different individuals.

3.3. Message Passing

The Graph Attention Network (GAT) [26] is a type of graph neural network, which can handle neighborhoods with varying scales while implicitly assigning individual weights to different nodes within a neighborhood. Under this mechanism, for any node $q \in \mathcal{N}_p^+$ in the graph, the output feature $\mathcal{H}_q^{(a)'}$ after a single layer of GAT with D attention mechanisms can be expressed as:

$$\mathcal{H}_q^{(a)'} = \bigoplus_{d=1}^D f_e(\alpha_{q,q}^d \mathbf{W}_s^d \mathcal{H}_q^{(a)} + \sum_{j \in \mathcal{N}(q)} \alpha_{q,j}^d \mathbf{W}_t^d \mathcal{H}_j^{(a)}), \quad (8)$$

where \bigoplus denotes the concatenation operation, and $f_e(\cdot)$ represents the Exponential Linear Unit (ELU) activation function [29], $\mathbf{W}_s^d \in \mathbb{R}^{K' \times K}$ and $\mathbf{W}_t^d \in \mathbb{R}^{K' \times K}$ are two learnable matrices in the d -th attention mechanism with K' being the output dimension, $\mathcal{N}(q) \in \mathcal{N}_p^+$ denotes the neighboring set that includes the nodes connecting with node q in the graph $\mathcal{G}_p^{(a)}$, and $\alpha_{q,j}^d$ represents the attention coefficient between nodes q and j , i.e.,

$$\alpha_{q,j}^d = f_l \left(f_l \left(\mathbf{a}_s^{d\top} \mathbf{W}_s^d \mathcal{H}_q^{(a)} + \mathbf{a}_t^{d\top} \mathbf{W}_t^d \mathcal{H}_j^{(a)} \right) + \mathbf{a}_e^{d\top} \mathbf{W}_e^d \mathcal{W}_{q,j}^{(a)} \right) \quad (9)$$

where $\mathbf{W}_e^d \in \mathbb{R}^{K' \times 1}$ is another weight matrix in the d -th attention mechanism, $\mathcal{W}_{q,j}^{(a)} \in \mathcal{W}_p^{(a)}$ is the assigned weight for the edge between nodes q and j , $\mathbf{a}_s^d \in \mathbb{R}^{K'}$, $\mathbf{a}_t^d \in \mathbb{R}^{K'}$ and $\mathbf{a}_e^d \in \mathbb{R}^{K'}$ are three learnable weight vectors, $f_l(\cdot)$ denotes the LeakyReLU activation function [30], and $f(\cdot)$ represents the Softmax function, which normalizes the computed attention coefficients to facilitate comparison of different nodes.

In our work, we cascade two GAT layers with $D_1 = 8$ and $D_2 = 1$ attention heads. The first layer encodes the node features into a tensor of dimension K' , which is then decoded in the second layer to recover the HRTF in the target node.

3.4. Dual GAT Structure

In Sec. 3.3, we manage to generate the HRTFs based on IM. However, according to [17], HRTFs also depend on some other factors, e.g., the absolute position of the target nodes themselves and the anthropometric characteristics. Therefore, we establish another graph where each vertex is determined by concatenating a and p , i.e., $\mathcal{M}_p^{(a)} = a \oplus p$, obtaining a new vertex set $\tilde{\mathcal{V}}_p^{(a)}$ as

$$\tilde{\mathcal{V}}_p^{(a)} = \{\mathcal{M}_q^{(a)} \mid q \in \mathcal{N}_p\} \cup \mathcal{M}_p^{(a)}. \quad (10)$$

Afterward, we combine (10) with the edge set (5) and the weight set (6), constructing a new graph $\tilde{\mathcal{G}}_p^{(a)}$ as

$$\tilde{\mathcal{G}}_p^{(a)} = \{\tilde{\mathcal{V}}_p^{(a)}, \mathcal{E}_p^{(a)}, \mathcal{W}_p^{(a)}\}. \quad (11)$$

According to (11), another GAT (i.e., the lower branch in Figure 2) is constructed to assist the upper GAT branch, which contains two GAT layers with $D_3 = 8$ and $D_4 = 1$ attention heads, respectively. Note that the output dimensions of the above two GAT layers are K' and K , which are identical to the upper branch described in Sec. 3.3.

The features are extracted from the target nodes in the output of two branches, which are then concatenated and fed into the fully connected layer, generating the predicted HRTF.

3.5. Loss Function

We adopt the Log-spectral distortion (LSD) as the loss function to guide network training, i.e.,

$$LSD(\mathcal{H}_p^{(a)}, \hat{\mathcal{H}}_p^{(a)}) = \sqrt{\frac{1}{K} \sum_{k=1}^K \left(20 \log_{10} \left| \frac{\mathcal{H}_p^{(a)}(k)}{\hat{\mathcal{H}}_p^{(a)}(k)} \right| \right)^2}, \quad (12)$$

where $\hat{\mathcal{H}}_p^{(a)}(k)$ represents the predicted HRTF of the k -th frequency bin.

4. EXPERIMENTS

4.1. Datasets

The experiments were carried out based on the HUTUBS dataset [28, 31], which includes HRIR data and anthropometric data from 96 subjects. These HRIR data were sampled on a spherical surface with a radius of 1.4 m, including measurement data from each subject at 440 positions and simulated data at 1730 positions. Each position yielded a 256-point HRIR signal for both left and right ears. The experiments used 12 features of the left ear and 13 features of the torso. Due to missing anthropometric data for three subjects (IDs 18, 79, and 92), the following experiments included data from 93 subjects.

4.2. Training

During the training stages, we employed a 5-fold cross-validation strategy to improve the generalization ability of the model. Two commonly used metrics, Log-spectral Distortion (LSD, the lower the better) and Spectral Distortion (SD, the lower the better), were adopted to evaluate the precision of HRTF predictions [25, 23]. SD was computed from the mean absolute error of the log-magnitude values between the estimated and true spectra. To further optimize the model, we employed the Adam optimizer [32], with an initial learning rate of 0.002 and a learning rate decay of 5% after each training epoch.

4.3. Comparison Results

In this part, we compared the D-GAT model ($\delta_\theta = 12$, $\gamma = 6.2$), the FiLM-HRTF model ($\delta_d = 0.2$) [25], and the linear interpolation method across different directions. The comparison was carried out on the horizontal plane (Hor.), the median plane (Med.), the frontal plane (Fro.) and arbitrary directions (Arb.). As shown in Table 1, the interpolation accuracy of all methods in Med. is higher compared to other planes. This is because the Med., being the symmetrical plane of the head, is generally less influenced by the human anatomical structure, and the distribution of sampling points on this plane is relatively uniform. In addition, the performance of the linear interpolation method is often limited due to the uneven distribution of sampling points in the HUTUBS dataset. The proposed D-GAT model outperforms other methods on all planes and directions, demonstrating its ability to effectively aggregate neighborhood information and reduce interpolation errors caused by uneven sampling distributions.

Table 1: The average LSD based on simulated data from the HUTUBS dataset.

Direction	Arb.	Hor.	Med.	Fro.
Linear Interp.	6.605	6.853	2.298	6.472
FiLM-HRTF [25]	1.768	2.199	1.108	2.116
D-GAT	1.527	2.008	1.050	1.814

To further evaluate the performance of the D-GAT, we also conducted a comparative experiment on a small-scale dataset (the real measurements from the HUTUBS dataset). The compared methods include the Linear Interp., the FiLM-HRTF model ($\delta_d = 0.3$), and the NIIRF model (using 32 peak filters) [24]. Due to the sparse nature of the measurement data, we selected relatively larger $\delta_\theta=24$ and $\gamma=12$ to ensure that each target point contains at least one neighboring point. Since NIIRF selects a fixed number of random azimuth and elevation angles to predict IIR filters to generate HRTFs, it is not applicable for comparison on different planes as in Table 1. Therefore, the comparison was carried out only in arbitrary directions. As shown in Table II, the D-GAT method performs the best, providing the smallest LSD (3.241 dB) and SD (1.924 dB) values. This further indicates the advantage of the D-GAT in handling sparse HRTF datasets in the real world.

Table 2: The average LSD and SD based on measured data from the HUTUBS dataset.

Method	LSD	SD
Linear Interp.	6.691	4.641
FiLM-HRTF [25]	3.449	2.156
NIIRF [24]	3.523	2.362
D-GAT	3.241	1.924

4.4. Ablation study

To evaluate the impact of different modules on overall performance, we performed ablation experiments on HUTUBS dataset. Table 3 shows the results when $\delta_\theta = 12$ deg. First, we adopted only the upper branch of D-GAT incorporating a complete graph (where $\gamma = \delta_\theta$) and all-one edge weights, named as the GAT+Complete Graph, to act as the baseline. Next,

by setting $\gamma=6.2$, we substituted the complete graph with a locally connected graph (LC-Graph) to form the GAT+LC-Graph, delivering a significant performance improvement in (b). The main reason is that the HRTFs are only strongly correlated between nearest nodes, and a GAT+LC-Graph structure allows the model to better learn such a relationship. Then, the edge weighting (6) was further adopted in (c), which achieves a better HRTF interpolation accuracy compared to (b). Finally, we added the second branch of GAT to carry out the DM from physical features to HRTFs, constructing the dual-branch GAT (D-GAT) with dual locally connected graphs (D-LC-Graph) in (d). It is clear that (d) performs better in Table 3, which demonstrates the effectiveness of the proposed D-GAT network architecture.

Table 3: Ablation study results, which averages the LSD from 5-fold cross-validation.

	Model	Graph Construction	LSD (dB)
(a)	GAT	Complete Graph	2.085
(b)	GAT	LC-Graph	1.746
(c)	GAT	LC-Graph + Edge Weights	1.567
(d)	D-GAT	D-LC-Graph + Edge Weights	1.523

4.5. Parameter Sensitivity Analysis

In this experiment, we evaluated the impact of parameters δ_θ and γ on HRTF interpolation using data from the fifth fold. The lower bound of γ was determined by the minimum distance between the nodes. As shown in Figure 3, LSD rarely changes with the variation of δ_θ or γ in a certain interval. To be specific, the largest fluctuation is about only 0.07 dB. This reveals that the proposed D-GAT model is relatively stable with the change of graph size or graph topology. In conclusion, the D-GAT is not too strict in the setting of the parameters δ_θ and γ .

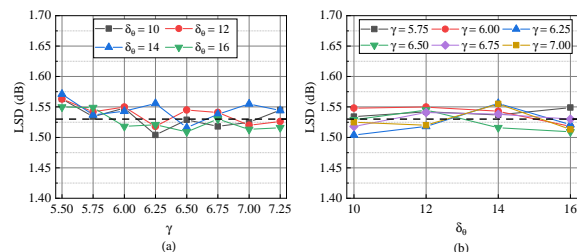


Figure 3: Performance evaluation under different parameters δ_θ and γ .

5. CONCLUSION

In this paper, we proposed a Dual-Graph Attention Network (D-GAT) for global HRTF interpolation. First, we formulated a set of graph construction rules, including neighborhood sampling, vertex (and edge) sets formulation, and edge weighting. Next, a dual-branch graph attention network was designed, in which one branch predicts the HRTF via reference HRTFs from neighboring nodes, while the other branch learns the relationship between the HRTF and physical features (i.e., the absolute target position and the anthropometric characteristics). The experimental results confirmed the validity of each module in the D-GAT. Compared with the state of the arts [24, 25], the D-GAT provided a higher HRTF interpolation accuracy.

6. Acknowledgement

This work was supported by the National Natural Science Foundation of China under Grants 62361045 and 62201297.

7. References

- [1] G. Ramos, M. Cobos, B. Bank, and J. A. Belloch, "A parallel approach to HRTF approximation and interpolation based on a parametric filter model," *IEEE Signal Processing Letters*, vol. 24, no. 10, pp. 1507–1511, Aug. 2017.
- [2] M. Johansson, "VR for your ears: Dynamic 3D audio is key to the immersive experience by mathias johansson · illustration by eddie guy," *IEEE Spectrum*, vol. 56, no. 2, pp. 24–29, Feb. 2019.
- [3] F. Zotter and M. Frank, *Ambisonics: A practical 3D audio theory for recording, studio production, sound reinforcement, and virtual reality*. Berlin, Germany: Springer Nature, 2019.
- [4] V. Sundareswaran, K. Wang, S. Chen, R. Behringer, J. McGee, C. Tam, and P. Zahorik, "3D audio augmented reality: implementation and experiments," in *IEEE/ACM Int. Symp. Mixed Augmented Reality*, Tokyo, Japan, 2003, pp. 296–297.
- [5] J. Yang and F. Mattern, "Audio augmented reality for human-object interactions," in *UbiComp/ISWC*, London United Kingdom, 2019, pp. 408–412.
- [6] Y.-C. Du, H.-C. Yu, W.-S. Ciou, and Y.-L. Li, "A wearable assistive listening device with immersive function using sensors fusion method for the 3D space perception," *IEEE Sensors Journal*, vol. 24, no. 2, pp. 2108–2117, 2023.
- [7] D. Vickers, M. Salorio-Corbetto, S. Driver, C. Rocca, Y. Levto, K. Sum, B. Parmar, G. Dritsakis, J. Albanell Flores, D. Jiang *et al.*, "Involving children and teenagers with bilateral cochlear implants in the design of the BEARS (both EARS) virtual reality training suite improves personalization," *Frontiers in Digital Health*, vol. 3, 2021.
- [8] S. Spagnol, "HRTF selection by anthropometric regression for improving horizontal localization accuracy," *IEEE Signal Processing Letters*, vol. 27, pp. 590–594, 2020.
- [9] 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.
- [10] C. Pörschmann and J. M. Arend, "Obtaining dense HRTF sets from sparse measurements in reverberant environments," in *AES 2019 Immersive Audio Conference*. New York, United States: Audio Engineering Society, 2019, pp. 1–10.
- [11] F. Grijalva, L. C. Martini, D. Florencio, and S. Goldenstein, "Interpolation of head-related transfer functions using manifold learning," *IEEE Signal Processing Letters*, vol. 24, no. 2, pp. 221–225, 2017.
- [12] K. Hartung, J. Braasch, and S. J. Sterbing, "Comparison of different methods for the interpolation of head-related transfer functions," in *AES International Conference*. Rovaniemi, Finland: Audio Engineering Society, 1999, pp. 319–329.
- [13] D. Poirier-Quinot and B. F. Katz, "The anaglyph binaural audio engine," in *AES Convention*, Milan, Italy, 2018.
- [14] 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, pp. 103–104, 2019.
- [15] J. M. Arend, F. Brinkmann, and C. Pörschmann, "Assessing spherical harmonics interpolation of time-aligned head-related transfer functions," *Journal of the Audio Engineering Society*, vol. 69, no. 1/2, pp. 104–117, 2021.
- [16] G. D. Romigh, D. S. Brungart, R. M. Stern, and B. D. Simpson, "Efficient real spherical harmonic representation of head-related transfer functions," *IEEE Journal of Selected Topics in Signal Processing*, vol. 9, no. 5, pp. 921–930, 2015.
- [17] A. O. 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*, vol. 32, pp. 2085–2099, 2024.
- [18] 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 *International Conference on Acoustics, Speech and Signal Processing*, Toronto, ON, Canada, 2021, pp. 3385–3389.
- [19] T.-Y. Chen, T.-H. Kuo, and T.-S. Chi, "Autoencoding HRTFs for DNN based HRTF personalization using anthropometric features," in *International Conference on Acoustics, Speech and Signal Processing*, Brighton, UK, 2019, pp. 271–275.
- [20] Y. Zhou, H. Jiang, and V. K. Ithapu, "On the predictability of HRTFs from ear shapes using deep networks," in *International Conference on Acoustics, Speech and Signal Processing*, Toronto, Ontario, Canada, 2021, pp. 441–445.
- [21] 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 *International Workshop on Acoustic Signal Enhancement*, Bamberg, Germany, 2022, pp. 1–5.
- [22] Y. Zhang, Y. Wang, and Z. Duan, "HRTF field: Unifying measured HRTF magnitude representation with neural fields," in *International Conference on Acoustics, Speech and Signal Processing*, Rhodes Island, Greece, 2023, pp. 1–5.
- [23] Y. Qiu, Z. Li, and J. Wang, "Individual HRTF prediction based on anthropometric data and multi-stage model," in *International Conference on Multimedia and Expo Workshops*, Brisbane, Australia, 2023, pp. 314–319.
- [24] 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 *International Conference on Acoustics, Speech and Signal Processing*, Seoul, Korea, 2024, pp. 1016–1020.
- [25] J. W. Lee, S. Lee, and K. Lee, "Global HRTF interpolation via learned affine transformation of hyper-conditioned features," in *International Conference on Acoustics, Speech and Signal Processing*, Rhodes Island, Greece, 2023, pp. 1–5.
- [26] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio, "Graph attention networks," *International Conference on Learning Representations*, p. 1–12, 2018.
- [27] V. R. Algazi, R. O. Duda, D. M. Thompson, and C. Avendano, "The CIPIC HRTF database," in *IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics*, New Platz, NY, USA, 2001, pp. 99–102.
- [28] B. Fabian, D. Manoj, P. Robert, W. J. Joschka, S. Fabian, V. Daniel, G. Peter, and W. Stefan, "The HUTUBS head-related transfer function (HRTF) database," 2019.
- [29] D.-A. Clevert, "Fast and accurate deep network learning by exponential linear units (elus)," *arXiv preprint arXiv:1511.07289*, 2015.
- [30] A. L. Maas, A. Y. Hannun, A. Y. Ng *et al.*, "Rectifier nonlinearities improve neural network acoustic models," in *International Conference on Machine Learning*, Atlanta, Georgia, USA, 2013.
- [31] F. Brinkmann, M. Dinakaran, R. Pelzer, P. Grosche, D. Voss, and S. Weinzierl, "A cross-evaluated database of measured and simulated HRTFs including 3D head meshes, anthropometric features, and headphone impulse responses," *Journal of the Audio Engineering Society*, vol. 67, no. 9, pp. 705–718, 2019.
- [32] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent.*, 2015.