ON THE PREDICTABILITY OF HRTFS FROM EAR SHAPES USING DEEP NETWORKS

Yaxuan Zhou, Hao Jiang, Vamsi Krishna Ithapu*

Facebook Reality Labs
yaxuanzh@uw.edu, haojiang@fb.com, ithapu@fb.com

ABSTRACT

Head-Related Transfer Function (HRTF) individualization is critical for immersive and realistic spatial audio rendering in augmented/virtual reality. Neither measurements nor simulations using 3D scans of head/ear are scalable for practical applications. More efficient machine learning approaches are being explored recently, to predict HRTFs from ear images or anthropometric features. However, it is not yet clear whether such models can provide an alternative for direct measurements or high-fidelity simulations. Here, we aim to address this question. Using 3D ear shapes as inputs, we explore the bounds of HRTF predictability using deep neural networks. To that end, we propose and evaluate two models, and identify the lowest achievable spectral distance error when predicting the true HRTF magnitude spectra.

Index Terms— AR/VR, spatial audio, 3D volume representation, 3D CNN, 3D residual U-Net

1. INTRODUCTION

Head-related impulse responses parameterize the transformations applied by the head (and ear) surface geometry on the acoustic signals as they enter the left and right ear canals. Their magnitude spectra are the head-related transfer functions (HRTFs), and along with interaural time differences, encode the information required for spatial audio perception. Since ear structures are unique to every individual, personalizing HRTFs is necessary for building robust/immersive augmented/virtual reality (AR/VR) systems. In principle, one can measure HRTFs acoustically, or synthesize them by numerical simulations on high-resolution 3D scans [1], for each individual. However, both these approaches are logistically and computationally expensive, and thus non-scalable for large populations of users – an obstacle to AR/VR applications. Alternatively, in the past decade, great efforts have been made to estimate HRTF individualization systems in a data-driven manner using machine learning approaches.

Typically HRTF prediction models use ear anthropometric features or ear images as inputs [2, 3, 4, 5], to either select an approximate HRTF from a database or synthesize a personalized HRTF using features of the target individual. Recently, deep networks were used to train a mapping from different ear representations to HRTF [6, 7, 8, 9]. Human-in-the-loop setups have also been proposed where subjects provide feedback during inference stage and the predicted HRTF is adjusted accordingly [10]. For a brief survey on HRTF personalization procedures, please refer to [1]. These prediction methods are constrained by the generality of the HRTF database, as well as whether HRTF and anthropometric features can be mapped to each other in some metric space.

Although several prediction models were proposed, many inherent problems are still not answered. First, the representational power of different ear inputs is not well understood. Most previous methods utilize the aforementioned ear-related inputs and euclidean loss functions, e.g. spectral distance error in log-magnitude domain. Specifically, anthropometric keypoints are usually selected empirically and labeled manually, and there is no compelling evidence yet that they can predict HRTF with high accuracy. Using 2D ear images for HRTF prediction also has limitations. The color images are often limited by viewing angles and self-occlusions. A recent study shows that some signal in HRTF cannot be extracted from anthropometric features or ear images [11].

Second, machine learning models (deep learning specifically) are data hungry, and current HRTF databases may be too small. Several research groups have built HRTF databases from acoustic measurements [12, 13, 14, 15, 16], with the largest database featuring 120 subjects [13]. Since the measurement setups are generally different, these may not be combined into a single uniform dataset. Alternatively, two databases used approx. 1000 synthetic ear shapes to simulate HRTFs [17, 18]. These are susceptible to domain discrepancy between synthesized ear shapes and actual human ear shapes (because the span of simulated ear shapes might be different from the span/manifold of HRTFs). Another recent open-sourced dataset of simulated HRTFs (from 3D scan of subjects) [19] also has small size (96 subjects).

Our main goal is to explore the limits of HRTF predictability with ear-related input representations. We use a larger dataset, and we build and evaluate deep neural networks (DNN) with 3D point cloud ear representations, thereby establishing a lower bound of HRTF estimation error

*The work was done while Yaxuan Zhou was a research intern at Facebook Reality Labs in Redmond, WA, USA.

with such highly informative (and costly) inputs. We propose and evaluate two models for this purpose.

2. HRTF PREDICTION FROM EAR SHAPES

2.1. Dataset & Data Representation

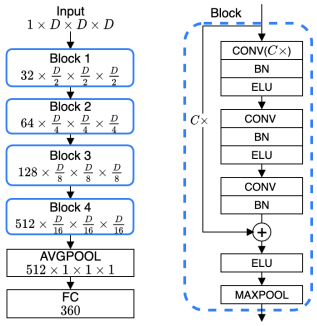
We utilize HRTFs from 645 human subjects for this study. Artec 3D scanners are used to obtain 3D meshes of head and upper torso for these subjects. Extensive quality checks with repeated measurements are performed to ensure $<$ 1mm mesh representational errors. The left-ear and right-ear HRIRs on a 1-meter sphere are simulated using the acquired 3D mesh with finite-difference time-domain (FDTD) method. Simulations were validated by comparing to acoustic measurements. The resulting far-field HRTF magnitude spectra are gammatone smoothed (ERB-filter) with 40 center frequencies located between 650Hz and 16KHz. Since low frequency content in HRTFs does not contain notches/peaks and simulations at high frequencies are noisy, we focus on the 30 frequencies in $1-12$KHz. HRTFs are simulated on a spatial grid with 10° resolution in both azimuth $\in [0^\circ, 360^\circ)$ and elevation $\in [-90^\circ, 90^\circ]$. However, without loss of generality, we restrict the evaluations to 10 elevations $\in [-30^\circ, 60^\circ]$.

The input left and right ear meshes are cropped and centered around the ear canal entrance. We voxelize them into volume data facilitating the DNN construction. Each mesh would correspond to a 3D tensor with equal size along each dimension. We use three different 3D tensor sizes: 16^3, 32^3 and 64^3, which correspond to 5.6mm, 2.8mm and 1.4mm per voxel respectively. [20] showed that scan precision of approx. 4mm is sufficient to maintain overall spectral shape of simulated HRTF (although 16^3 setting is just above this threshold, we use it to investigate models' representative capacity). The left ear-HRTF pair and right ear-HRTF pair are used as separate data points. We mirror all right ear-HRTF pairs to form new left ear-HRTF pairs, which increases the dataset size to a total of 1290 ear-HRTFs pairs for evaluations.

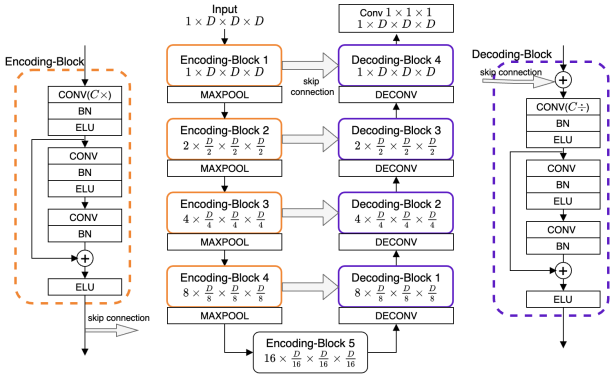
2.2. DNN Models

HRTF prediction is a regression problem. Recall from Section 2.1 that the HRTF has 360 directions and 30 frequency bins, and our input is 3D tensor. We propose two different DNN architectures that map the 3D ear tensor to the corresponding HRTF parameterized by 30×360 magnitude values. The networks are illustrated in Figure 1. Both models allow for joint predictions of HRTFs across directions, and they differ in terms of the hidden representational space.

CNN-Reg: This network comprises of a cascade of blocks, each of which contains convolution layers, batch normalization and ELU non-linearity layers as shown in Figure 1(a). Output from the last such block is fed into an adaptive average pooling layer, followed by a fully connected layer. This last layer is a major bottleneck for network size; and so instead of jointly predicting all the 30 frequency bins and all 360 directions, we train 30 separate networks, one per frequency bin. While reducing the network size, this allows us to check the influence of frequency-dependent simulations errors on the model performance (as we will discuss later in Section 3).



(a) Network architecture of CNN-Reg



(b) Network architecture of UNet-Reg

Fig. 1. Network architecture of (**a**) CNN-Reg and (**b**) UNet-Reg. D: voxel dimension; C×/C÷: increasing/reducing channels; K: kernel size; S: stride; P: padding; CONV: 3D conv layer with K=3, S=1, P=1; BN: batch norm; ELU: ELU activation; AVGPOOL: adaptive average pooling; MAXPOOL: 3D maxpooling with K=2, S=2, P=0; FC: fully connected; DECONV: transpose conv layer with K=3, S=2, P=1.

UNet-Reg: Recall that HRTFs are inherently spherical, i.e., they can be represented as a 3D volume across directions. Existing works on classifying spherical inputs like [21, 22], although partly relevant, do not directly apply for our problem of regression over spherical domain outputs (HRTF) from volumetric inputs (ear shape). Hence, as an alternative to CNN-Reg, we propose 3D UNet inspired by [23, 24] to map 3D inputs to 3D outputs. Similar to CNN-Reg, we also train one UNet-Reg per frequency. Figure 1(b) shows the network architecture, and it captures several unique aspects of HRTF prediction. First, observe that unlike the conventional UNet, there is no direct voxel-to-voxel correspondence between inputs and outputs as the HRTF values only exist on a spherical

surface. Thus we define a spherical surface in the 3D output tensor, and from that sphere, we pick 360 values with the selected azimuths/elevations as stated in Section 2.1 to represent HRTF values at 360 directions. This design also allows for scaling up HRTF prediction on denser spatial grids without increase in computational overhead. We can also define multiple smaller concentric spheres in the 3D output tensor to include near-field HRTF predictions. Second, the spatial upsampling operation in UNet captures the hypothesis that ears and HRTFs can jointly be modeled as over-parameterized hidden representations (possibly the intrinsic manifold on which the spatial audio signals reside). This also implies that the information sharing across directions in UNet-Reg vs. CNN-Reg would be different.

Spectral distance error (SDE), defined below, is used as the loss function for backpropagation. $\hat{h}(\theta, \varphi, f)$ and $h(\theta, \varphi, f)$ are the predicted and ground-truth (simulated) HRTF magnitudes at azimuth θ, elevation φ and a frequency bin centered at f. N_d is the total number of directions.

$$\texttt{SDE}(f) = \frac{1}{N_d} \sum_{\theta,\varphi} \left| 20 \log \frac{\hat{h}(\theta, \varphi, f)}{h(\theta, \varphi, f)} \right| \quad (1)$$

3. EXPERIMENTS

CNN-Reg and UNet-Reg are evaluated using 1290 HRTF-ear pairs (800/200/290 for training, validation and testing) and a 5-fold cross validation was done. Both left and mirrored-right HRTF-ear pairs from a given subject are included in only one of training, validation or testing to avoid any overlap. For optimization, we used Adam with variable learning rate (0.1, decaying eventually to 0.0001), batch size of 30, and 100 epochs. Exhaustive validation of hyperparameters is done and we are reporting the best outcomes. We use two baselines: a simulated HRTF on KEMAR dummy head (denoted by **genHRTF**), and a population average of HRTFs from across training/validation sets (denoted by **pop-avg**). Since overly simple models underfit and predict average output for all inputs, we use pop-avg to evaluate the efficacy of models in learning useful information. We also compare CNN-Reg and UNet-Reg with existing HRTF prediction pipelines. All the models are implemented in PyTorch.

CNN-Reg & UNet-Reg vs. pop-avg & genHRTF: Figure 2 summarizes the frequency-dependent SDE (with mean and s.d. across directions and subjects in top and bottom plots respectively). Overall, our proposed methods significantly outperform the genHRTF at all frequencies, clearly asserting the need for individualized prediction of HRTFs. In lower frequencies, CNN-Reg, UNet-Reg predict as well as pop-avg. This makes sense because, for such large wavelengths, individual differences in fine-grained ear structures do not affect the HRTF spectrum estimation, and so the models are learning to predict an average. In mid-range frequencies (e.g., $3 - 12$kHz), CNN-Reg and UNet-Reg significantly outperform pop-avg by at least 1dB, indicating that they are extracting useful features from the input space compared to the population average. CNN-Reg slightly outperforms UNet-Reg. However, UNet-Reg has several advantages. Firstly, UNet-Reg has fewer network parameters (35k vs. 17m for CNN-Reg) leading to a smaller footprint during inference. Secondly, UNet-Reg's volumetric output allows for a more intuitive/interpretable network design, especially for prediction over denser spatial grid of directions (including near field).

SDE vs. frequencies & directions: Observe that, in Figure 2, the SDE deviation over subjects (bottom) is larger than deviation over directions (top), indicating that HRTF signals have higher variance across subjects, than across directions for a given subject. Also, SDE deviation across subjects increases at higher frequencies (bottom) while the deviation across directions decreases (top), implying the simulation error at high frequencies may be subject-specific.

To further understand the frequency dependence, we first noticed from ground truth that magnitudes at higher frequencies have larger variance (across subjects) than those at lower frequencies. This is also seen in SDE trends which generally increase (almost monotonically) from low to high frequencies for all models. Second, we also observe that the ground truth HRTF magnitudes on ipsilateral directions have smaller variance than those on contralateral directions (which in our case are the areas around azimuth = 270°). Even more so, the variance of HRTF magnitude almost always monotonously decreases as elevation angle increases from bottom to the top. Following this variance trend, Figure 3 shows that CNN-Reg performs better in ipsilateral, above-horizontal regions and between $1.5 - 7$kHz. To eliminate this influence of chang-

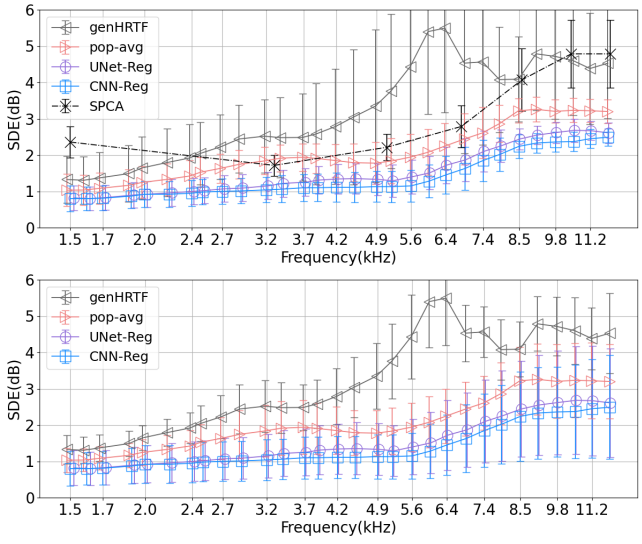


Fig. 2. Mean and s.d of SDE vs. frequencies. (**Top**) s.d across directions; (**Bottom**) s.d across subjects.

ing variance across different frequencies/directions on model performance, we normalized HRTFs to have zero mean and unit variance on each frequency and direction and retrained the network; but found no change in performance. Hence we claim that the high SDE in the contralateral and below-horizon directions may be mostly caused by non-ear-related structures like chin or upper torso, which the model won't be able to learn from the ear shape input. And the high SDE at higher frequencies is most likely caused by simulation error.

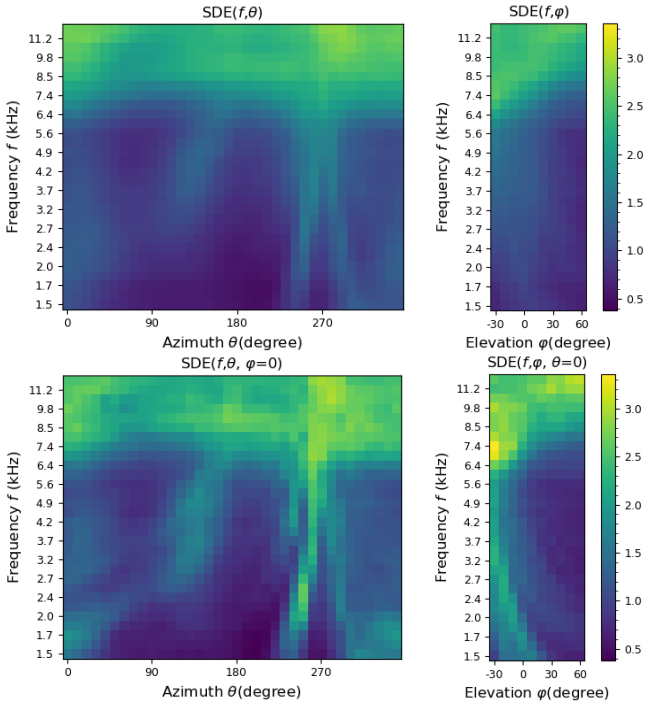


Fig. 3. SDE of CNN-Reg (with 32^3 grid) averaged over elevation (**Top left**) and azimuth (**Top right**). **Bottom left, bottom right** are SDEs at elevation $0°$ and azimuth $0°$, respectively.

Are ear scans better for predicting HRTFs? As motivated in Section 1, one of our goals is to establish predictability bounds of HRTFs from high-resolution inputs like 3D ear scans. To do this, we report our performance against the proposals in Chen19 [9] and Zhang20 [11]. Zhang20 uses a spatial principal component analysis (SPCA) for learning HRTF basis embeddings, while Chen19 uses an encoder-decoder for embedding HRTFs, and a DNN to predict the embedding. Both use anthropometric features as inputs. Figure 2 (top) shows Zhang20 with respect to CNN-Reg and UNet-Reg; clearly, HRTFs are better predicted via voxelized meshes. Table 1 shows the mean SDE across frequencies. Zhang20's SDE is averaged over an overlapping set of frequencies; and we use the same (see Figure 2 for range), while Chen19's SDE is averaged across 25 azimuths at $0°$ elevation and 173 frequencies within the range of $200 - 15$kHz. The datasets in these studies and ours are different, however, trends in Table 1 indicate improvement in SDE achieved by using highly-informative inputs.

Effect of voxelization: Table 2 shows the influence of input sizes on performance. The mean SDE is averaged over all frequencies and directions (s.d is across directions). CNN-Reg and UNet-Reg achieve best SDEs (1.38 ± 0.38dB and 1.52 ± 0.41dB) with 32^3 and 64^3 voxel grids respectively. Smaller grid of 16^3 increases error. Since overparameterization may result in better network optimization, increasing CNN-Reg's size may reduce jump in error from 32^3 to 64^3 voxel grid; however this is not desirable because network footprint increases rapidly. Since SDEs are generally the same across changing input sizes, we can claim that CNN-Reg and UNet-Reg learn useful information for HRTF prediction, independent of these voxel grid sizes.

Table 1. SDEs of ear scans vs. anthropometric features. Note: Zhang20 and Chen19 use different (smaller) dataset.

	CNN-Reg	**UNet-Reg**	**Zhang20***	**Chen19***	**genHRTF**
SDE	**1.67**	1.84	3.24	3.43	3.63

Table 2. Mean SDE vs. input size (s.d across directions)

Input Grid	$16 \times 16 \times 16$	$32 \times 32 \times 32$	$64 \times 64 \times 64$
CNN-Reg	1.49 ± 0.36	$\mathbf{1.38 \pm 0.38}$	1.57 ± 0.43
UNet-Reg	1.61 ± 0.45	1.53 ± 0.38	$\mathbf{1.52 \pm 0.41}$

Comparison with numerical simulations: We asked in Section 1 whether a DNN can replace the numerical simulator. While accuracy was one factor, the per-subject inference time is equally critical for scalable HRTF individualization. While numerical simulation takes 20-30 mins per subject, models' inference takes tens of millisecs. CNN-Reg and UNet-Reg both offer computationally effective alternatives to simulations with an average tolerance of 1.38dB and 1.52dB SDE in prediction. We do note that simulations themselves are shown to have errors at high frequencies [25], and so some of this error might be due to the noise in ground truth data, rather than model's incapacity in estimation. This also relates to increasing SDE vs. frequency in Figures 2 and 3. Also note that, our simulated HRTFs utilize meshes of torso and head while our proposed methods only take ear shape as input; possibly leading to the 'floor' of performance around 1dB in these plots.

4. CONCLUSION

We proposed two DNN models that predict HRTFs from 3D ear tensors. We achieved highest prediction accuracy yet; showing lower bounds of achievable errors using highly informative ear shape inputs. Future work includes using perceptual loss functions and improving model design.

5. REFERENCES

[1] C. Guezenoc and R. Seguier, "Hrtf individualization: A survey," *arXiv preprint arXiv:2003.06183*, 2020.

[2] D. N. Zotkin, J. Hwang, R. Duraiswaini, and L.S. Davis, "Hrtf personalization using anthropometric measurements," in *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2003, pp. 157–160.

[3] X. Liu and X. Zhong, "An improved anthropometry-based customization method of individual head-related transfer functions," in *ICASSP*, 2016, pp. 336–339.

[4] P. Bilinski, J. Ahrens, M. R. P. Thomas, I. J. Tashev, and J. C. Platt, "Hrtf magnitude synthesis via sparse representation of anthropometric features," in *ICASSP*, 2014, pp. 4468–4472.

[5] J. He, W. Gan, and E. Tan, "On the preprocessing and postprocessing of hrtf individualization based on sparse representation of anthropometric features," in *ICASSP*, 2015, pp. 639–643.

[6] C. J. Chun, J. M. Moon, G. W. Lee, N. K. Kim, and H. K. Kim, "Deep neural network based hrtf personalization using anthropometric measurements," in *143rd Audio Engineering Society International Convention*, 2017, p. 9860.

[7] H. Fayek, L. van der Maaten, G. Romigh, and R. Mehra, "On data-driven approaches to head-related-transfer function personalization," in *143rd Audio Engineering Society International Convention*, 2017, pp. 1–10.

[8] G. Lee and H. Kim, "Personalized hrtf modeling based on deep neural network using anthropometric measurements and images of the ear," *Applied Sciences*, vol. 8, pp. 2180, 2018.

[9] T. Chen, T. Kuo, and T. Chi, "Autoencoding hrtfs for dnn based hrtf personalization using anthropometric features," in *ICASSP*, 2019, pp. 271–275.

[10] K. Yamamoto and T. Igarashi, "Fully perceptual-based 3d spatial sound individualization with an adaptive variational autoencoder," *ACM Transactions on Graphics (TOG)*, vol. 36, no. 6, pp. 1–13, 2017.

[11] 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.

[12] V. R. Algazi, R. O. Duda, D. M. Thompson, and C. Avendano, "The cipic hrtf database," in *Proceedings of the 2001 IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics*, 2001, pp. 99–102.

[13] P. Majdak, M. J. Goupell, and B. Laback, "3d localization of virtual sound sources: Effects of visual environment, pointing method, and training," *Attention, Perception & Psychophysics*, vol. 72, pp. 454–469, 2010.

[14] K. Watanabe, Y. Iwaya, Y. Suzuki, S. Takane, and S. Sato, "Dataset of head-related transfer functions measured with a circular loudspeaker array," *Acoustical Science and Technology*, vol. 35, pp. 159–165, 2014.

[15] T. Carpentier, H. Bahu, M. Noisternig, and O. Warusfel, "Measurement of a head-related transfer function database with high spatial resolution," in *Proceedings of the 7th Forum Acusticum, European Acoustics Association*, 2014.

[16] R. Bomhardt, M. de la Fuente Klein, and J. Fels, "A high-resolution head-related transfer function and three-dimensional ear model database," in *Proceedings of the 172nd Meeting of the Acoustical Society of America*, 2016, p. 050002.

[17] S. Ghorbal, X. Bonjour, and R. Séguier., "Computed hrirs and ears database for acoustic research," in *Audio Engineering Society Conference: 148th International Conference*, 2020.

[18] C. Guezenoc and R. Séguier, "A wide dataset of ear shapes and pinna-related transfer functions generated by random ear drawings," *The Journal of the Acoustical Society of America*, vol. 147, pp. 4087–4096, 2020.

[19] 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, pp. 705–718, 2019.

[20] 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 *ICASSP*, 2018, pp. 551–555.

[21] T. S. Cohen, M. Geiger, J. Köhler, and M. Welling, "Spherical cnns," in *ICLR*, 2018.

[22] N. Perraudin, M. Defferrard, T. Kacprzak, and R. Sgier, "Deepsphere: Efficient spherical convolutional neural network with healpix sampling for cosmological applications," *Astronomy and Computing*, vol. 27, pp. 130–146, 2019.

[23] O. Cicek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3d u-net: Learning dense volumetric segmentation from sparse annotation," 2016.

[24] K. Lee, J. Zung, P. Li, V. Jain, and H. S. Seung, "Superhuman accuracy on the snemi3d connectomics challenge," 2017.

[25] S. T. Prepelita, J. Gomez, M. Geronazzo, R. Mehra, and L. Savioja, "Pinna-related transfer functions and lossless wave equation using finite-difference methods: Verification and asymptotic solution," *The Journal of the Acoustical Society of America*, vol. 146, no. 5, 2019.