DOI: 10.61782/fa.2023.0795

COMPARISON OF DEEP-NEURAL-NETWORK ARCHITECTURES FOR THE PREDICTION OF HEAD-RELATED TRANSFER FUNCTIONS USING A PARAMETRIC PINNA MODEL

Florian Pausch* **Felix Perfler** **Nicki Holighaus** **Piotr Majdak**

Acoustics Research Institute, Austrian Academy of Sciences, 1040 Vienna, Austria

ABSTRACT

Personalised head-related transfer functions (HRTFs) represent a key component for applications in virtual or augmented reality with high demands in perceptual audio. Numerical methods enable the calculation of personalised HRTFs based on the individual geometry of a listener. Such a geometry can be generated by exploiting the information from multi-view-plus-depth (MVPD) images of the listener's pinna. However, the results are typically noisy, especially in regions relevant for HRTF calculations. To address this shortcoming, HRTFs can also be calculated based on a mesh obtained from a parametric pinna model (PPM) whose parameters were optimised to fit noisy measurements of an individual ear geometry. The feasibility of this approach has been investigated by employing a convolutional neural network (CNN) which predicts the PPMs parameters from synthetic MVPDs pinna-only images. In this contribution, we varied the type of network architecture and analysed the effects on prediction accuracy. The results of comparative evaluations will be discussed in the geometric domain to explore the limits of the PPM-based HRTF personalisation, and corroborate the choice of a feasible network architecture before implementing further optimisation.

Keywords: *personalised binaural audio, ear modelling, machine learning, regression.*

1. INTRODUCTION

Personalised audio is more than ever in demand both by industrial and academic research to achieve a high level of plausibility when binaurally reproducing virtual acoustic environments. Taking advantage of the rapid process in the field of machine learning, deep neural networks (DNNs) have been extensively applied for the personalisation of HRTFs [1, 2]. Such DNNs can, for example, be trained to directly predict the log-magnitude spectrum of HRTFs from edge-detected single-view ear images and individual anthropometric features [3]. Despite the application of sophisticated network architectures across studies, the HRTFs predicted this way often suffer from inaccurately restored monaural cues and may feature direction-dependent errors.

The availability of a PPM [4] combined with methods to numerically calculate HRTFs from meshes [5] opens up new possibilities for HRTF personalisation. We propose an indirect approach in which the parameters of said PPM are predicted by a DNN from MVPD images for the best possible adaptation of a target-ear geometry. The use of MVPD images allows the DNN extracting latent information from concave pinna regions which may be occluded in single-view images, and helps mitigate estimation errors provided images that were captured at varying and unknown camera distances. The optimised PPM instance represents an approximation of the original target mesh and, once stitched to a head mesh, enables numerically calculating HRTFs for arbitrary directions using software packages like COMSOL [6] or MESH2HRTF [7]. As additional benefit, such a model-based approach also does not involve problems emerging from artefacts that are typically contained in photogrammetrically obtained target-ear geometries.

**Corresponding author*: *florian.pausch@oeaw.ac.at.*

Representing a subset of DNNs, CNNs with residual layers [3, 8] have demonstrated their feasibility for image-classification or object-detection tasks by utilising a sliding window to learn local features, and skip connections to facilitate feature reusability. The architectural category of Vision Transformers widely surpass the performance of CNNs in such tasks by utilising self-attention layers to assess the global relevance of local features within a window [9]. One shortcoming is their increased resource requirements when applied to high-resolution images. Also, they only use image patches with fixed size hindering to extract variable amounts of relevant information within a window. To address these issues, the shifted-window (Swin)-Transformer architecture was introduced [10, 11]. Increased efficiency is achieved by applying a shifted-window multi-head self-attention strategy in which local features are learned by merging initially smaller and increasingly larger image patches within local windows to construct hierarchical feature maps. The use of shifted windows also enables to share local features across adjoining windows.

Since it is unclear which one of these two fundamentally different neural-network architectures is best suited when applied to our regression task of estimating optimised PPM parameters, they are subject to evaluation. In the current study, both training and inference of each architecture are based on a synthetic MVPD dataset which we generated using the PPM. The accuracy in predicted PPM instances is evaluated in the geometric domain to facilitate an informed choice of a suitable architecture. No analysis in the acoustic domain is carried out, rendering the current work a preliminary study on the feasibility of the proposed indirect HRTF personalisation approach.

2. METHODS

The proposed concept for indirect HRTF personalisation is presented below by describing the main functional elements, including the PPM, its application for the creation of synthetic MVPD pinna-only images, and the two DNN network architetures evaluated in this study.

2.1 Parametric pinna model

The PPM used in this study is defined in BLENDER [12] and described in detail by Pollack *et al.* [4]. Summarised compactly, the model consists of a well-defined template mesh, representing a generic adult human ear, which is connected to an armature modifiable via 144 parameters, These PPM parameters allow controlling the global pinna translation and rotation, and anisotropically scaling the whole pinna mesh. They further facilitate deformation of local pinna regions by means of a distance-based automated-weighting function. The local deformation is accomplished by rotating or translating the start and end points of the underlying Bézier curves, as well as isotropically scaling their intermediate curve segments. Shape keys form a special subset of these PPM parameters and are available for the refinement of concave pinna regions and anthropometric peculiarities.

2.2 Dataset

In this work, we only varied the local location parameters of the PPM [4]. Assuming mutually independent continuous uniform PPM-parameter distributions, each PPM-parameter set was drawn from multiple independently sampled PPM-parameter distributions over the empirically set interval $[-1, 1]$ mm around the default values. Note that this seemingly small variation of individual PPM parameters in combination already leads to a substantial deformation of the template mesh [4], see Figure 4. In total, 10 000 PPM instances were generated. We used PYTHON (v3.10.8) and the module `bpy` (BLENDER PYTHON API, v3.5.0) to modify the PPM. Per PPM instance, MVPD pinna-only images, i.e. `png` images and `OpenEXR` depth images, were rendered at a resolution of 256×256 px using BLENDER's Cycles engine for 25 camera perspectives. The camera was positioned equidistantly to the center of the ear-canal entrance at the ipsilateral ear side within the azimuth and elevation-angle intervals of $[225°, 315°]$ and $[-50°, 50°]$, which were discretised in steps of 22.5° and 25°, respectively. For each perspective, the camera rotation was set to point at the ear-canal entrance. Depth information was linearly mapped between the origin of the world coordinate system (black) and the current camera position (white) to normalised units. We subsequently split the data into 72 % training images, 8 % images reserved for intermediate validation, and 20 % test images.

2.3 Concept

Figure 1 presents how the created dataset of MVPD images can be used within the scope of the proposed indirect HRTF-personalisation concept based on a DNN. During training, the DNN aims at learning features from the MVPD training images to update the DNN weights

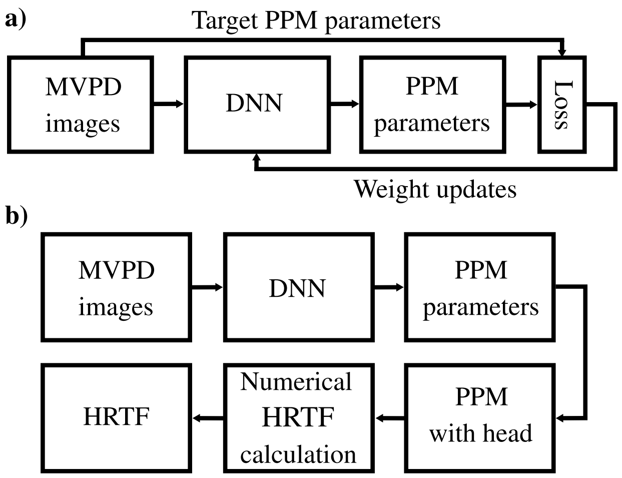

Figure 1: Concept for indirect HRTF personalisation. **a)** Training of a DNN to predict the PPM parameters from MVPD images by minimising a specific loss function and correspondingly updating the DNN weights. **b)** Application of the learned DNN weights during inference to predict the PPM parameters from unknown MVPD images and create an optimised PPM instance for subsequent numerical HRTF calculation.

with the aim to minimise the loss between target and predicted PPM parameters. For inference, the learned optimised weights are applied to the DNN to predict the PPM parameters for each MVPD test image and create the corresponding PPM instance. This optimised PPM instance can subsequently be used to numerically calculate HRTFs for arbitrary directions.

2.4 Network architectures

The proposed indirect HRTF-personalisation concept was evaluated with two state-of-the-art DNN architectures.

The first architecture used is a modified version of CNN-Reg [2], representing a CNN with skip connections. This architecture features four consecutive stages and was originally applied to predict the single-frequency gammatone-filtered HRTF log-amplitude spectrum jointly for 360 directions from voxelised ear meshes. In our study, we reduced the input dimensionality to account for MVPD images, and jointly estimated all 54 local location parameters in the fully-connected output layer.

Table 1: Specific implementation of the compared DNN architectures.

	Architecture	
Configuration	CNN-Reg	Swin-XT
#Stages	4	3
#Output channels per stage	[32, 64, 128, 512]	
#Heads per stage		[32, 64, 128]
#Blocks per stage	[1, 1, 1, 1]	[2, 3, 2]
Kernel size	3	
Window size		8

As second architecture, we used a scaled-down Swin Transformer [13]. Prior work has demonstrated that Swin Transformers are well-suited for the extraction of features, for example, from 3D images at different levels of resolution [14]. Transferred to the current problem, this architecture is expected to be useful for efficiently extracting locally restricted pinna features and weighting their relevance for improved learning. After observing that already the tiny variant of the Swin Transformer, Swin-T [11, 13], is prone to overfitting most likely due to being too large for our problem, we further reduced its complexity. The resulting extra-tiny variant is referred to as Swin-XT.

Both DNN architectures were implemented in PyTorch (v2.0.0) and their differing configuration settings are listed in Table 1. For a fair comparison, we aimed at matching the number of DNN weights to approximately 5.8 M, which resulted from adopting CNN-Reg to the current problem. In both architectures, the number of input channels was set according to the number of camera perspectives plus the corresponding depth images, i.e. to 50. As optimiser and loss function, we used Adam with default PyTorch settings and the Huber loss, respectively. Both models were trained for 800 epochs using a batch size of 32. A cosine-annealing schedule with the maximum number of iterations set to the maximum number of epochs was applied. For inference and per architecture, the learned weights from the epoch exhibiting the lowest validation loss were loaded to test the two architectures using the MVPD images rendered based on the corresponding PPM test instances.

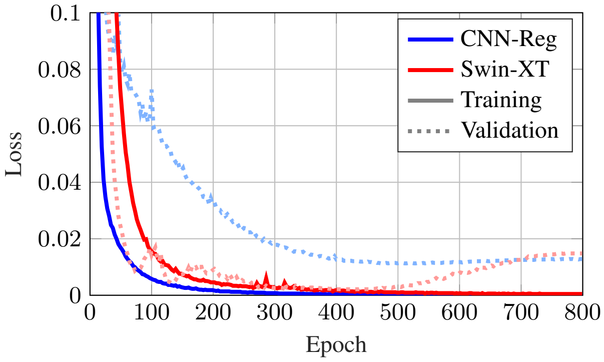

Figure 2: Loss progressions of CNN-Reg (blue lines) and Swin-XT (red lines) during training (solid lines) and validation (dotted lines).

2.5 Error metric

We considered the Pompeiu–Hausdorff distance [15]

$$d_{\mathrm{H}} = \max\left\{\sup_{a\in A} d(a,B),\ \sup_{b\in B} d(b,A)\right\} \tag{1}$$

as geometrical error metric to describe the similarity between two point clouds with $d(\cdot)$, $\sup(\cdot)$, and A, $B \subset \mathbb{R}^3$ representing the directed point-wise distance, the supremum, and the optimised PPM instance and corresponding target mesh, respectively. To obtain an average performance metric, we additionally evaluated the mean and median values of the point-wise distances for the direction with the larger supremum.

3. RESULTS AND DISCUSSION

Figure 2 displays the training and validation losses of both architectures. Although both training losses show similar exponential decays, the validation loss of CNN-Reg takes substantially longer to converge to its minimum of approximately 0.011 at epoch 513. This minimum is about 5.4 times higher than the one achieved by the Swin-XT architecture which amounts to approximately 0.002 at epoch 429. While the validation loss of CNN-Reg remains almost constant after convergence, it increases in Swin-XT after having reached its minimum, which indicates overfitting. A likely better ability to generalise to unknown MVPD images, particularly at the loss minimum, can be attributed to the Swin-XT architecture. This assumption is deduced from the better corresponding training and validation losses of the Swin-XT, compared to the ones obtained for CNN-Reg whose validation loss is deviating from and consistently higher than its training loss.

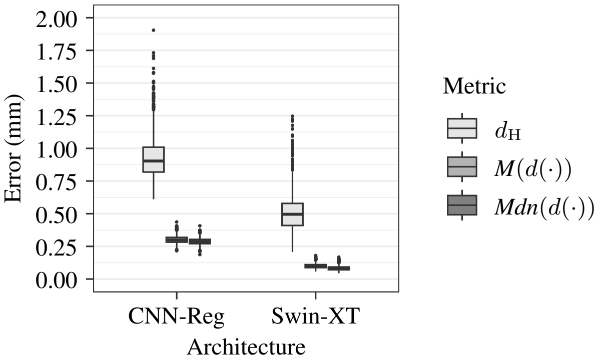

Figure 3: Geometric error metrics averaged over the PPM test instances. Boxplots show the Pompeiu-Hausdorff distance d_{H}, and median (*Mdn*) and mean (*M*) values of the pointwise minimum distance $d(\cdot)$ for the direction with larger supremum.

To quantify these performance differences in terms of geometric errors, Figure 3 presents the Pompeiu-Hausdorff distance d_{H}, cf. Equation (1), averaged over all PPM test instances being parametrised as predicted by CNN-Reg and Swin-XT. In addition, the mean (*M*) and median (*Mdn*) values of the point-wise minimum distances [4] for the direction containing the larger supremum are shown. Being in line with the expectations after having assessed the loss progressions, CNN-Reg resulted in higher errors ($M \pm SD$) compared to Swin-XT in terms of d_{H} (0.93 ± 0.15 vs. 0.5 ± 0.14), $M(d(\cdot))$ (0.3 ± 0.03 vs. 0.1 ± 0.02), and $Mdn(d(\cdot))$ (0.29 ± 0.03 vs. 0.08 ± 0.02). The results of three independent-sample t-tests, conducted between each data pair per metric type, indicated that the differences in means are statistically significant, $p < .001$, suggesting a generally better performance of Swin-XT. For both architectures, the error magnitudes are substantially smaller than the ones reported in previous studies when manually fitting the same PPM to photogrammetrically obtained target ears [4]. However, those target ears required adaptations of all types of global and local PPM parameters, and shape keys. Also, they did not match the PPM in number of points and with regard to its basis shape. Thus, a PPM registration involving limited parameter variations and "matched" meshes, as in the current study, generally

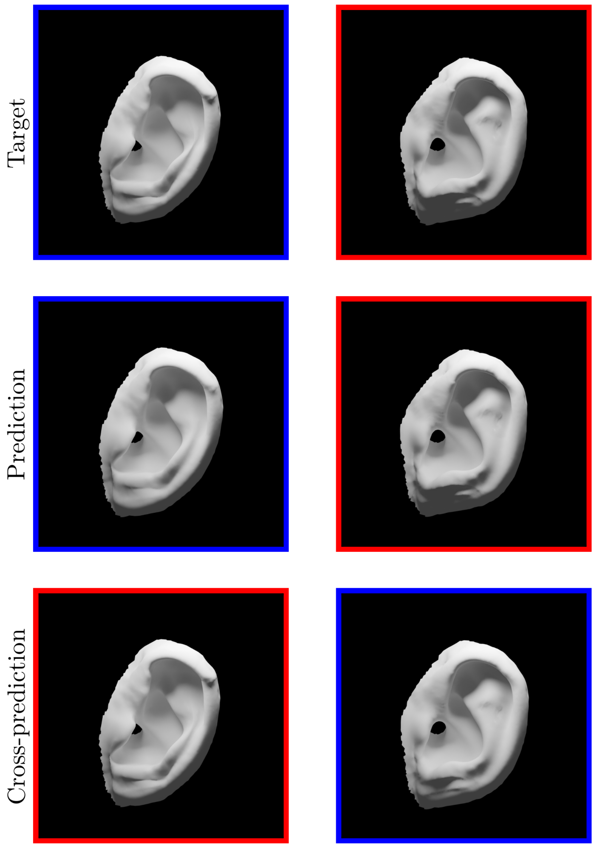

Figure 4: Renderings of the "worst" PPM test instances as predicted (second row) and cross-predicted (third row) by CNN-Reg (blue frames) and Swin-XT (red frames) in comparison to the respective target ears (first row).

represents a somewhat less challenging scenario. The fact that the median values of d_{H}, $M(d(\cdot))$ and $Mdn(d(\cdot))$ in both architectures are below $1\,\mathrm{mm}$ given combined PPM-parameter variations indicates a mesh representation being likely accurate enough for the calculation of personalised HRTFs that perceptually will not differ from their target equivalents [5].

Finally, the worst-case predictions and their rendering results were examined. For the determination of the prediction with the lowest average accuracy per architecture, i.e. the "worst" PPM test instance, we selected the PPM test instance with the largest $Mdn(d(\cdot))$. Figure 4 visualises the rendering results. The first row presents the rendered PPM target ears with color-coded frames to indicate the affiliation to each architecture. The second row shows the rendered PPMs as predicted by each architecture. In the third row, the cross-prediction of one architecture is shown provided the same MVPD images of the PPM target instance presented to the respective other architecture. Although the worst-case predictions using CNN-Reg is generally quite accurate, it lacks a detailed rendition of the pinna fine structure in comparison to the worst-case prediction of Swin-XT. This shortcoming can also be spotted in the cross predictions. Only the Swin-XT architecture is capable of rendering even the finest pinna details, which visually confirms the particularly low validation loss, see Figure 2, and geometric errors, see Figure 3. In comparison, CNN-Reg results in spatially smoothed representations of the target ears, which may partially be attributed to its reduced generalisation capability, see Figure 2.

Overall, these results support the superiority of the Swin-XT architecture and its effectiveness in applying local attention mechanisms. However, it remains to be shown to what extent the residual geometric errors impact the objective and perceptual quality of the correspondingly calculated HRTFs.

4. CONCLUSION

We presented a novel concept for indirect HRTF personalisation in which a DNN is applied for the prediction of an optimised parameter set from MVPD images to parametrise a PPM which can be used to numerically calculate personalised HRTFs. Two fundamentally different DNN network architectures, CNN-Reg and Swin-XT, with a matched number of weights were trained based on synthetic MVPD images obtained from randomly parametrised PPM instances, featuring variations of local location parameters only. A set of unknown synthetic MVPD images was used to test the trained DNNs. In comparison to CNN-Reg, the results of the evaluation in the geometric domain revealed that the Swin-XT architecture showed superior performance and was able to recreate pinna details of unknown target ears more accurately. It is therefore considered better suited for predicting the local location parameters of the PPM.

5. ACKNOWLEDGMENTS

This study received funding from the European Union's Horizon 2020 research and innovation funding programme (grant no. 101017743, project "SONICOM"– Transforming auditory-based social interaction and communication in AR/VR).

6. REFERENCES

[1] R. Miccini and S. Spagnol, "HRTF Individualization using Deep Learning," in *2020 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, Mar. 2020, pp. 390–395.

[2] Y. Zhou, H. Jiang, and V. K. Ithapu, "On the Predictability of HRTFs from Ear Shapes Using Deep Networks," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 441–445.

[3] G. W. Lee and H. K. Kim, "Personalized HRTF Modeling Based on Deep Neural Network Using Anthropometric Measurements and Images of the Ear," *Applied Sciences*, vol. 8, no. 11, 2018. [Online]. Available: `https://www.mdpi.com/2076-3417/8/11/2180`.

[4] K. Pollack, F. Pausch, and P. Majdak, "Parametric pinna model for a realistic representation of listener-specific pinna geometry," in *24th International Congress on Acoustics: ICA 2022*, Invited paper., Gyeongju, Korea: The Acoustical Society of Korea, Oct. 24, 2022. [Online]. Available: `https://ica2022korea.org/data/Proceedings_A21.pdf`.

[5] H. Ziegelwanger, P. Majdak, and W. Kreuzer, "Numerical calculation of head-related transfer functions and sound localization: Microphone model and mesh discretization," *The Journal of the Acoustical Society of America*, vol. 138, pp. 208–222, 2015.

[6] COMSOL AB, *COMSOL Multiphysics*, 2023. [Online]. Available: `https://www.comsol.com/`.

[7] H. Ziegelwanger, W. Kreuzer, and P. Majdak, "Mesh2HRTF: An open-source software package for the numerical calculation of head-related transfer functions," Florence, Italy, 2015, p. 42 583.

[8] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *Computer Vision – ECCV 2016*, Cham: Springer International Publishing, 2016, pp. 630–645, ISBN: 978-3-319-46493-0.

[9] A. Vaswani *et al.*, "Attention is all you need," in *Advances in Neural Information Processing Systems*, I. Guyon *et al.*, Eds., vol. 30, Curran Associates, Inc., 2017.

[10] Z. Liu *et al.*, "Swin transformer v2: Scaling up capacity and resolution," 2021. [Online]. Available: `https://arxiv.org/abs/2111.09883`.

[11] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, *A ConvNet for the 2020s*, 2022. [Online]. Available: `https://arxiv.org/abs/2201.03545`.

[12] Blender Development Team, *Blender - a 3D modelling and rendering package*, 2023. [Online]. Available: `http://www.blender.org`.

[13] Z. Liu *et al.*, "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, Los Alamitos, CA, USA: IEEE Computer Society, Oct. 2021, pp. 9992–10 002. [Online]. Available: `https://doi.ieeecomputersociety.org/10.1109/ICCV48922.2021.00986`.

[14] A. Hatamizadeh, V. Nath, Y. Tang, D. Yang, H. R. Roth, and D. Xu, "Swin UNETR: Swin Transformers for Semantic Segmentation of Brain Tumors in MRI Images," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, Cham: Springer International Publishing, 2022, pp. 272–284, ISBN: 978-3-031-08999-2.

[15] D. Pompeiu, "Sur la continuité des fonctions de variables complexes," fr, *Annales de la Faculté des sciences de l'Université de Toulouse pour les sciences mathématiques et les sciences physiques*, vol. 2e série, 7, no. 3, pp. 265–315, 1905.