JASA EXPRESS LETTERS

Correspondence between three-dimensional ear depth information derived from two-dimensional images and magnetic resonance imaging: Use of a neural-network model

Cite as: JASA Express Lett. **1**, 112001 (2021); https://doi.org/10.1121/10.0007151
Submitted: 09 September 2021 • Accepted: 22 October 2021 • Published Online: 11 November 2021

Tiernan Watson, Joe Halse, Granit M. Dula, et al.

openaccess

View Online

Export Citation

ARTICLES YOU MAY BE INTERESTED IN

Chinese Abstracts
Chinese Journal of Chemical Physics **34**, i (2021); https://doi.org/10.1063/1674-0068/34/04/cabs

Momentum conservation in the Biot–Savart law
American Journal of Physics **89**, 1033 (2021); https://doi.org/10.1119/10.0005207

Bell's theorem, non-computability and conformal cyclic cosmology: A top-down approach to quantum gravity
AVS Quantum Science **3**, 040801 (2021); https://doi.org/10.1116/5.0060680

ASA

© 2021 Author(s).

JASA EXPRESS LETTERS ARTICLE asa.scitation.org/journal/jel

Correspondence between three-dimensional ear depth information derived from two-dimensional images and magnetic resonance imaging: Use of a neural-network model

Tiernan Watson, Joe Halse, Granit M. Dula, Noorpreet Soni, Yue Wu, and Ifat Yasin[a)]
Department of Computer Science, University College London, 66-72 Gower Street, London WC1E 6BT, United Kingdom
tiernan.watson.17@ucl.ac.uk; Joe.Halse.17@ucl.ac.uk; granit.dula.17@ucl.ac.uk; noorpreet.soni.17@ucl.ac.uk; yue.wu.17@ucl.ac.uk; i.yasin@ucl.ac.uk

Abstract: There is much interest in anthropometric-derived head-related transfer functions (HRTFs) for simulating audio for virtual-reality systems. Three-dimensional (3D) anthropometric measures can be measured directly from individuals, or indirectly simulated from two-dimensional (2D) pinna images. The latter often requires additional pinna, head and/or torso measures. This study investigated accuracy with which 3D depth information can be obtained solely from 2D pinna images using an unsupervised monocular-depth estimation neural-network model. Output was compared to depth information obtained from corresponding magnetic resonance imaging (MRI) head scans (ground truth). Results show that 3D depth estimates obtained from 2D pinna images corresponded closely with MRI head-scan depth values. *© 2021 Author(s). All article content, except where otherwise noted, is licensed under a Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).*

[Editor: Charles C. Church] https://doi.org/10.1121/10.0007151

Received: 9 September 2021 **Accepted:** 22 October 2021 **Published Online:** 11 November 2021

1. Introduction

Head related transfer functions (HRTF) capture the temporal and spectral scattering of audio waves by the ears, head, and torso of a listener as it travels from an external sound source to the ear canal. Since it is impractical to measure personalised HRTFs for users of augmented reality/virtual reality (AR/VR) systems, generic HRTFs (measured using a manikin or directly from individuals and averaged across a subpopulation) are often used. Generic HRTFs may provide sub-optimal auditory localisation cues for the user, so alternative methods can be used to model/synthesise a close approximation to personalised HRTFS. Such methods include using (a) frequency scaling (Middlebrooks and Green, 1992), (b) selection of appropriate HRTFS from a database (Barumerli *et al.*, 2018), and (c) a linear mapping from the principle component analysis (PCA) weights applied to users' anthropometric parameters to PCA weights applied to the HRTF spectrum (Lee and Kim, 2018; Meng *et al.*, 2018). Some of these approaches may provide reasonable approximations for only a limited range of spatial locations, require onerous anthropometric measures, or are too computationally demanding to be of routine and widespread use.

There is currently great interest in using anthropometric-based estimates of HRTFs (Chun *et al.*, 2017; Zhu *et al.*, 2017; Brinkmann *et al.*, 2019; Islam and Tashev, 2020). HRTFs can be estimated based on deep neural networks using direct measurements from the individual (pinna, head, torso, etc.) (Bilinski *et al.*, 2014; He *et al.*, 2015), which are then combined with 2D images of the ear (Lee and Kim, 2018). A more parsimonious approach would be one in which the requirement for direct anthropometric data from the individual is reduced such that a sole 2D image would suffice to provide sufficient information for selection of the appropriate HRTF.

Depth maps can be obtained using monocular, binocular, or multiview vision. These methods describe the level of independence in the control of the viewing device (or eye). Monocular refers to the situation where each viewing device (or eye) is used separately, and binocular refers to the situation where both viewing devices are used together (Lauer *et al.*, 2011). For a variety of binocular and multi-view based methods, stereo vision technology is commonly used (Saxena *et al.*, 2008). Generally, stereo vision technology involves simultaneously taking photos of a subject from two angles. Epipolar geometry is then used to find the corresponding pair of points on the photos of the subject, from which the disparity map (which stores the corresponding pixel shift between two photos from different angles) is obtained (Zhang, 2018). Depth values can then be obtained from the disparity map.

[a)] Author to whom correspondence should be addressed.

Some earlier methods required semi-manual methods; user input was required to mark visual elements in the foreground or background (Phan *et al.*, 2011). Methods can also use a deep learning approach (requires training on pairs of input and desired output images) such that the network learns to infer the output image from only the input image. On the other hand, convolutional neural networks (CNN) are a common type of deep network that require only an input and the ground truth image for training (Saxena *et al.*, 2008). Generally, specialist equipment is required to acquire the training data. This could be a depth camera or scanning laser which can capture ground truth simultaneously with an aligned photograph. Examples of such databases are the KITTI dataset [compiled using a scanning laser distance sensor mounted on a vehicle, alongside cameras creating images of outdoor environments such as roads, buildings, and trees (Godard *et al.*, 2017; Song and Kim, 2019)] and the "Make3D" dataset [consisting of 1000 outdoor scenes and 50 interior views captured with a scanning laser and corresponding photograph (Mo *et al.*, 2013)].

With regards to inferring HRTFs from 2D pinna images, in the first step a 2D input image can be obtained of the user's pinnae, (taken by a camera). In the second step the user has to input additional anthropometric data relating to the pinnae through a graphical user interface (Zotkin *et al.*, 2003). The input data is then combined with a low-frequency "head-and-torso" model to derive the relevant HRTF.

The current project investigates the feasibility of bypassing the requirement for a user to input any additional data by using an unsupervised monocular-depth estimation neural-network model to extract depth information from 2D pinna images, to investigate whether a one-step approach may be sufficient. To assess sufficiency, the obtained depth information is compared to the ground truth [i.e., depth information obtained from magnetic resonance imaging (MRI) head scans associated with the pinna images].

2. Methods

2.1 Data

2D images of pinnae were obtained from the SADIE II database (SADIE, 2019) which provides a broad range of ears under consistent light settings. This database was used as the 2D pinna images are also associated with the respective 3D MRI head scans of the individuals; this allows us to perform a comparison of the depth information retrieved from the 2D images and the actual (ground truth) depth information provided by the MRI scans. The SADIE II database holds ear images for twenty individuals, for both the left and right ear. The zip file per individual contains the 2D pinna images of both ears (with black grid lines overlaid on the image), and the associated 3D head scans.

2.2 Image processing

In order to generate depth maps from the 2D images, unsupervised monocular depth estimation neural network models were used (Godard *et al.*, 2017). Godard *et al.* (2017) outlined the architecture of a fully convolutional model that does not require any depth data, but is instead trained to synthesize depth as an intermediate, i.e., the neural network learns to perform single image depth estimation in the absence of ground truth depth data. This is achieved within the model by implementation of a novel training loss that enforces left-right depth consistency within the network. The training loss module and the approach to use left images to produce disparities for both images (left and right) is described in Figs. 2 and 3 of Godard *et al.* (2017).

The models were trained for 50 epochs on images with a resolution of 512×256 on a batch size of 8. The models generated the depth map needed to transform the 2D image of the pinna to a 3D representation, but it does so as a colour map, where the colours change from blue to orange to bright yellow as the depth, from the plane of the view of the image, decreases. This gives a visualisation of how well the algorithm has detected the change in depth across the structures of the pinna, which allows for qualitative assessment of the results. The models provided by Godard *et al.* (2017) were used on 20 pinna images provided by the SADIE II database in order to extract the depth estimates of the pinna. Briefly, the images were pre-processed, after which a CNN model was applied to generate a colour depth map. The 2D pinna images were originally overlaid with black grid lines. First an image pre-processing program was written in MATLAB based on the use of a median filter to process the image [threshold-based salt and pepper noise removal (Hereford and Rhodes, 1988)], to reduce the effect of the pixels associated with the black lines. Figure 1 shows a pinna image before and after algorithm application.

In this way the black grid lines overlying the pinna images of 36 individual ears (left and right pairs from the SADIE II dataset) were removed. In order to further improve the quality of the generated depth maps from these images, the brightness, contrast, and sharpness of the images were consistently adjusted. The optimum image parameters were: brightness −10, contrast +100, and sharpness 230 (Photoshop's units of measurement).

2.3 model training and application

The depth estimation algorithm was applied to each processed image with each of the algorithm's six different models, which had been trained on different datasets. This was undertaken in order to compare the effectiveness of the different models. Figures 2–4 shows three different pinna images and comparison of the processed pinna image and the resulting depth map produced using the KITTI trained model. The CNN model was also run on all the images to generate all the

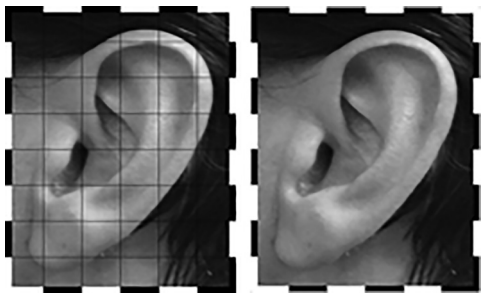

Fig. 1. A side by side comparison of one of the original pinna images with grid lines (left) and a new image without grid lines (right) after algorithm application.

colored depth images. The resultant depth image maps the colors ranging from yellow to orange through to purple to black, in order of increasing depth distance. The points listed in the tables below are also shown marked in the corresponding pinna images (corrected for resolution differences between the depth dataset and pinna image).

2.4 Depth maps extracted from the associated MRI scans

To compare the generated depth values from the pinna images to the ground truth, depth values also needed to be obtained from the head scan data (in a compatible format). An OPENGL program was written in C++ that computes depth in the fragment shader and maps it from 0 to 1 using the centre of the head model as the greatest depth (black). The vertex shader was a standard projection, view, model transformation for the rendering, but a different variable is used so the model is transformed into camera (eye) space and then the z co-ordinate is used to calculate the distance. Compared to ray casting, this option is less computationally expensive. Once the head scan image was created, the raw values could be written to csv file and the greyscale depth image produced. As there was no information about how far the pictured ears were to the camera, we used the closest vertex of the ear as the zero-depth point.

2.5 Depth comparison between 2D image and MRI scan

For each 3D pinna depth image the numerical depth values per pixel could be compared to the numerical value of the respective ground truth. For the latter, comparisons were made using an estimate centroid. Tables 1, 2, and 3 present the depth data for the pinna image corresponding to the images in Figs. 2, 3, and 4 (with and without grid lines) after processing and the equivalent ground truth. The pixel points chosen for the tables were selected to show data from equivalent positions per ear shape. The pixel points were chosen not randomly, but on similar areas of the ear to provide good comparisons. Generally, on inspection the image processing enhancements made an improvement to the accuracy of the estimated depth image.

3. Results

A comparison of the pinna image depth estimated using the algorithm (as described above) and the ground truth (the depth values estimated from the head scans), show that in general, there is a good correspondence between the depth values estimated from the 2D pinna images and the head scan depth values. Removal of the gridlines also made a substantial difference as expected. For instance, analysing the results of image H6, the total absolute error between the depth values obtained from the "with grid lines" image and the depth values obtained from the head scans: $|11.49 - 10.74| + |13.84 - 10.88| + |9.75 - 11.24| + |8.99 - 10.70| + |11.74 - 10.74| + |10.35 - 11.09| = 8.65$ cm.

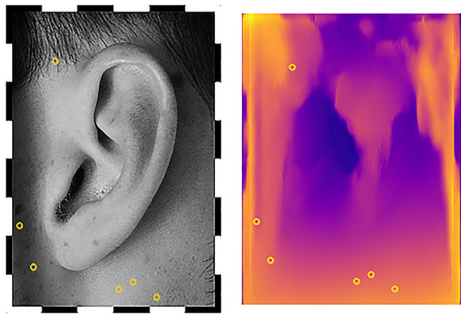

Fig. 2. A side by side comparison of one of the processed pinna images (H6) after applying the median filter algorithm and image enhancement (left) and the resultant depth image (right).

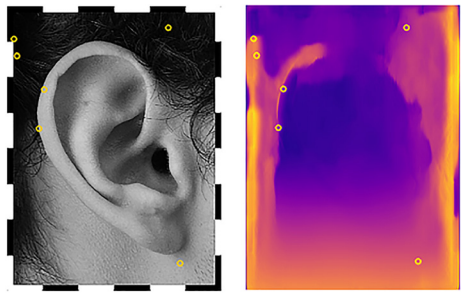

Fig. 3. Comparison layout as for Fig. 2 for pinna image H9.

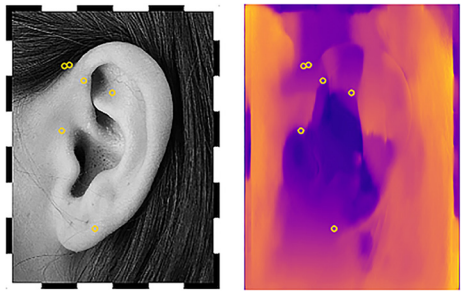

Fig. 4. Comparison layout as for Fig. 2 for pinna image H20.

Table 1. Depth values for select pixel points for each step of the image processing of the H6 ear image from the SADIE II database. The bold values are the values that have the smallest absolute error to the ground truth value. Depth values are presented for images with gridlines, without gridlines, and the processed image. Depth values were also derived from the head scans and are presented under the heading of "ground truth."

Pixel point	With grid (cm)	Without grid (cm)	Processed (cm)	Ground truth (cm)
(267, 235)	11.49	**10.61**	11.57	10.74
(117, 47)	13.84	9.51	**11.88**	10.88
(67, 217)	9.75	12.69	**11.51**	11.24
(34, 183)	8.99	**11.84**	12.2	10.7
(301, 229)	11.74	**11.52**	11.83	10.74
(357, 242)	10.35	10.53	10.52	11.09

Table 2. As for Table 1, for pinna image H9.

Pixel point	With grid (cm)	Without grid (cm)	Processed (cm)	Ground truth (cm)
(77, 110)	26.4	10.7	**11.24**	11.37
(25, 45)	4.89	**7.23**	6.47	11.8
(384, 20)	7.5	**11.23**	11.12	11.33
(411, 230)	**10.94**	7.72	7.41	9.56
(18, 30)	5.09	**7.76**	7.01	11.84
(90, 75)	9.52	**12.61**	12.47	12.82

Table 3. As for Table 1, for pinna image H20.

Pixel point	With grid (cm)	Without grid (cm)	Processed (cm)	Ground truth (cm)
(254, 79)	7.71	8.72	**11.16**	10.74
(186, 68)	13.36	9.47	**11.47**	10.57
(140, 55)	12.03	11.76	**11.53**	10.74
(134, 113)	11.56	**10.58**	11.85	10.68
(213, 201)	14.15	10.25	**10.35**	10.68
(152, 54)	13.11	11.52	**11.79**	11.72

Performing a similar calculation for the depth values obtained from the without grid lines image and the processed image gives 5.43 and 5.26 cm, respectively. For the latter estimations, this is a 37.2% and 39.1% reduction in absolute error respectively. In addition, applying the image processing enhancements also had an impact on improving the quality (a 3.1% error reduction) in this case.

Similarly, for H9, the absolute error between depth values obtained from the with grid lines image and head scans is 33.47 cm, whereas the absolute error between the depth values obtained from the without grid lines image and head scans is 15.2 cm which is a 54.6% reduction in error. The absolute error between depth values obtained from the processed image and the head scans is 13 cm which is a 61.2% reduction in error compared to the grid line version and a 14.4% reduction in error compared to the processed image. In general, the absolute difference between the depth estimates obtained from the 2D pinna images (after gridlines and processing) and the depth values obtained from the head scans differed by very little: 0.88, 2.6, and 0.74 cm for H6, H9, and H20 respectively.

4. Discussion and conclusion

This study evaluated the extent to which depth data could be extracted from 2D pinna images (using an unsupervised monocular-depth estimation neural-network model) and the degree of correspondence with the ground truth (depth values obtained from the corresponding head scans).

The advantage of a method which can accurately predict depth from only 2D images of the pinna is that the user does not need to provide additional anthropometric data. The formation and shape of the ear can actually very noticeably be seen from the colour depth images, in particular regions of the outer pinna ridge and contour of the concha. This demonstrates that the model performs relatively well in mapping and estimating the image depth values, including finer pinna contours. However, there is also a lot of noise produced, as would be expected. This is most likely a result of the large contrast between the hairline and/or hair colour, and the skin. Other approaches which could further improve the mapping of voxel to pixel, in this particular application, are ones that utilize a fuzzy clustering approach [e.g., Zhang *et al.* (2020)]. There are relatively few neural network model algorithms trained for close-up images (such as pinnae, as in this study) to evaluate correspondence between extracted depth information and ground truth. However, there is one example which also suggests good correspondence between estimated depth and ground truth: Nicodemou *et al.* (2018) inferred depth maps from photographs of human hands trained with ground truth data captured with a corresponding aligned Kinect sensor, and demonstrated close correspondence to the ground truth. An alternative approach may be to look towards using models such as Godard *et al.* (2019) (which use a multiscale sampling method) or Johnston and Carneiro (2020) (which uses discrete disparity prediction) to further reduce visual artefacts. The current approach used a model architecture trained on the KITTI database. Further studies may adopt a more targeted approach, using a model trained with a subsets of pinna images, or salient aspects of the input, such as to improve the subsequent analyses.

Other approaches for deriving HRTFS from images require extensive time for training and have not specifically tested depth correspondence between input pinna image and inferred depth (albeit indirectly from derived HRTF or ground-truth depth information). For instance, a neural network [CNN and/or deep neural network (DNN)] can be trained on data from the CIPIC database to derive HRTFs (Hu *et al.*, 2008; Lee and Kim, 2018). However, such networks require a considerable database for training purposes and it remains unknown to what extent the derived HRTF would have differed from the individual's measured HRTF. Furthermore, recent studies suggest that some areas of the pinna may have more of an effect on the HRTF, such as the concha or areas close to the triangular fossa (Stitt and Katz, 2021); subsequent studies may look towards focussing on depth extraction from these areas.

In order to improve the quality of the colour images that a model produces, the input images should resemble the input the model expects to receive. In the current study, the ambient lighting of the input images, as well as content differed from the ones that the KITTI model was trained on (the KITTI model was trained using car traffic images). Additionally, for the model to be effective, especially a neural network model, it needs a large training dataset for training. One way to possibly overcome this issue is to have the images produced on a large scale. This could be achieved by an ear image generating neural network model, similar to the image generating model DRAW (Gregor *et al.*, 2015). This kind of model would allow a large variety of ear images to be produced relatively quickly. Of course, it would be simpler to apply the model to images without gridlines, but gridlines are often applied in order to match salient pinna structures across pinnae and to the MRI head scans. In addition, the issue of extracting the pixels relevant to the ear in the colour depth image could also be addressed by using a machine vision library like OpenCV (Boyko *et al.*, 2018) to identify those pixels. In conclusion, the CNN model described in Godard *et al.* (2017) can be used to derive 3D depth information from a 2D photo of the pinna, which corresponds well with the ground truth depth data obtained directly from the head scans of the same individuals. Such a method would allow for a speeded correlation and selection of the appropriate HRTF from a pinna photo without the need for further additional anthropometric data to be provided.

References and links

Barumerli, R., Geronazzo, M., and Avanzini, F. (**2018**). "Round Robin comparison of inter-Laboratory HRTF measurements—Assessment with an auditory model for elevation," in IEEE 4th VR *Workshop Sonic Interactions Virtual Environments (SIVE)*, pp. 1–5.

Bilinski, P., Ahrens, J., Thomas, M. R. P., Tashev, I. J., and Platt, J. C. (**2014**). "HRTF magnitude synthesis via sparse representation of anthropometric features," In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4468–4472.
Boyko, N., Basystiuk, O., and Shakhovska, N. (**2018**). "Performance evaluation and comparison of software for face recognition, based on dlib and OpenCV library," In *2018 IEEE Second International Conference on Data Stream Mining & Processing (DSMP)*, c 478–482.
Brinkmann, F., Dinakaran, M., Pelzer, R., Grosche, P., Voss, D., and Weinzierl, S. (**2019**). "A cross-evaluated database of measured and simulated HRTFs including 3D head meshes, anthropometric features, and headphone impulse responses," J. Audio Eng. Soc. **67**, 705–718.
Chun, C. J., Moon, J. M., Lee, G. W., Kim, N. K., and Kim, H. K. (**2017**). "Deep neural network based HRTF personalization using anthropometric measurements," *AES, 143rd Convention*, New York, NY.
Godard, C., Mac, A. O., and Brostow, G. J. (**2017**). "Unsupervised monocular depth estimation with left-right consistency," In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 270–279.
Godard, C., Mac, A. O., Firman, M., and Brostow, G. J. (**2019**). "Digging into self-supervised monocular depth estimation," In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3828–3838.
Gregor, K., Danihelka, I., Graves, A., Rezende, D. J., and Wierstra, D. (**2015**). "Draw: A recurrent neural network for image generation," preprint arXiv:1502.04623.
He, J., Gan, W.-S., and Tan, E.-L. (**2015**). "On the preprocessing and postprocessing of HRTF individualization based on sparse representation of anthropometric features," In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 639–643.
Hereford, J. M., and Rhodes, W. T. (**1988**). "Nonlinear optical median filtering by time-sequential threshold decomposition," In *Hybrid Image and Signal Processing 939*, International Society for Optics and Photonics, pp. 40–47.
Hu, H., Zhou, L., Ma, H., and Wu, Z. (**2008**). "HRTF personalization based on artificial neural network in individual virtual auditory space," Appl. Acoust. **69**, 163–172.
Islam, M. T., and Tashev, I. (**2020**). "Anthropometric features estimation using integrated sensors on a headphone for HRTF personalization," In *Audio Engineering Society Conference on Audio for Virtual and Augmented Reality*.
Johnston, A., and Carneiro, G. (**2020**). "Self-supervised monocular trained depth estimation using self-attention and discrete disparity volume," In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4756–4765.
Lauer, M., Schonbein, M., Lange, S., and Welker, S. (**2011**). "3D object tracking with a mixed omnidirectional stereo camera system," Mechatronics **21**, 390–398.
Lee, G. W., and Kim, H. K. (**2018**). "Personalized HRTF modeling based on deep neural network using anthropometric measurements and images of the ear," Appl. Sci. **8**, 2180.
Meng, L., Wang, X., Chen, W., Ai, C., and Hu, R. (**2018**). "Individualization of head related transfer functions based on radial basis function neural network," In *IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6.
Middlebrooks, J. C., and Green, D. M. (**1992**). "Observations on a principal components analysis of head-related transfer functions," J. Acoust. Soc. Am. **92**, 597–599.
Mo, Y., Liu, T., Zhu, X., Dai, X., and Luo, J. (**2013**). "Segment based depth extraction approach for monocular image with linear perspective," In *International Conference on Intelligent Science and Big Data Engineering*, pp. 168–175.
Nicodemou, V. C., Oikonomidis, I., Tzimiropoulos, G., and Argyros, A. (**2018**). "Learning to infer the depth map of a hand from its color image," arXiv:1812.02486.
Phan, R., Rzeszutek, R., and Androutsos, D. (**2011**). "Semiautomatic 2D to 3D image conversion using scale-space random walks and a graph cuts based depth prior," in *Proceedings of the 18th IEEE International Conference on Image Processing*, pp. 865–868.
Saxena, A., Chung, S. H., and Ng, A. Y. (**2008**). "3D depth reconstruction from a single still image," Int. J. Comp Vision **76**, 53–69.
Song, M., and Kim, W. (**2019**). "Depth estimation from a single image using guided deep network," IEEE Access **7**, 142595–142606.
Spatial Audio for Domestic Interactive Entertainment (SADIE) (**2019**). https://www.york.ac.uk/sadie-project/database.html.
Stitt, P., and Katz, B. F. (**2021**). "Sensitivity analysis of pinna morphology on head-related transfer functions simulated via a parametric pinna model," J. Acoust. Soc. Am. **149**, 2559–2572.
Zhang, S. (**2018**). "High-speed 3D shape measurement with structured light methods: A review," Opt. Lasers Eng. **106**, 119–131.
Zhang, Y., Govindaraj, V., Murugan, P. R., and Sankaran, S. (**2020**). "Smart identification of topographically variant anomalies in brain magnetic resonance imaging using a fish school based fuzzy clustering approach," in *IEEE Transactions on Fuzzy Systems*.
Zhu, M., Shahnawaz, M., Tubaro, S., and Sarti, A. (**2017**). "HRTF personalization based on weighted sparse representation of anthropometric features," In *IEEE International Conference on 3D Immersion (IC3D)*, pp. 1–7.
Zotkin, D. Y. N., Hwang, J., Duraiswaini, R., and Davis, L. S. (**2003**). "HRTF personalization using anthropometric measurements," In *Proceedings of IEEE Workshop Applications of Signal Processing to Audio and Acoustics*, pp. 157–160.