DOI: 10.61782/fa.2025.0244

FORUM ACUSTICUM
EURONOISE 2025

COMPLEX ROOM GEOMETRY INFERENCE VIA ACOUSTIC ECHOES

Inmo Yeon[1] **Jung-Woo Choi**[1*]
[1] School of Electrical Engineering, KAIST, South Korea

ABSTRACT

Estimation of room geometry is crucial for realistic audio rendering in virtual and augmented reality, as well as for applications like sound field reconstruction. This study introduces a deep learning-based method that infers room geometry by directly predicting floorplan and height maps from room impulse responses (RIRs). Unlike traditional approaches that estimate room parameters such as wall positions and room size, this segmentation-based approach predicts a detailed geometric floorplan map of a room, allowing it to handle irregular and complex shapes, including curved walls. By utilizing high-order reflections, the proposed method captures complex geometric details, even those unobservable from the position of the audio device due to occlusion, which are challenging to resolve with conventional methods relying on first-order reflections. The model's exploitability of high-order reflections is demonstrated through gradient activation map visualizations and experiments with RIRs limited to first-order reflections, highlighting their critical role in reconstructing complex geometries. Validated on synthetic datasets, including Manhattan and Atlanta layouts, the model demonstrates high accuracy in reconstructing diverse room geometries, exhibiting robustness in scenarios with indoor furniture and objects.

Keywords: Room geometry inference, room impulse response, deep neural network

1. INTRODUCTION

Room geometry inference (RGI) is essential for various audio applications, including immersive virtual and augmented reality (VR/AR) experiences, source separation, and sound field reconstruction. Accurate room geometry information can help simulate realistic room impulse responses and thus enable immersive audio rendering [1]. Also, room geometry information can improve the source separation and enhancement performance [2]. Although vision-based RGI approaches using a panoramic image of the indoor scene have shown effectiveness [3,4], they struggle with non-line-of-sight (NLOS) walls that are invisible due to occlusion by other walls. Moreover, acoustically meaningful geometric information should be captured for immersive audio rendering, highlighting the need for acoustic-based RGI methods.

Acoustic-based RGI methods utilize room impulse responses (RIRs) to extract time-of-arrival (TOA) information. Previous human-curated methods for acoustic-based RGI mainly utilize TOAs of first-order reflections, which provide distance to the walls [5–9]. Recently, learning-based methods applying deep neural networks (DNNs) have been proposed to overcome the constraints of conventional human-curated methods, which generally rely on the estimation of room or planar wall parameters using TOAs of first-order reflections [10–12]. Although these DNN-based techniques have shown promising RGI performance, they cannot be applied to complex rooms with curved walls. In this paper, we introduce our recent work [13] tackling this challenge by reformulating the RGI task as a pixel-level segmentation problem. This approach allows us to estimate general room geometry with an arbitrary number of walls and more general wall shapes including curved ones.

2. PROPOSED METHOD

In this work, we consider a 3D room geometry $\mathbf{Y}^{\mathrm{3D}} \in \mathbb{R}^{b\times b\times h}$, whose floor and ceiling are parallel. Such room geometries can be decomposed into a 2D floorplan map $\mathbf{Y}^{\mathrm{LW}} \in \mathbb{R}^{b\times b}$ sampled by b pixels for length-width space and a 1D height map $\mathbf{y}^{\mathrm{H}} \in \mathbb{R}^{h}$ sampled by h pixels for

*Corresponding author: jwoo@kaist.ac.kr.

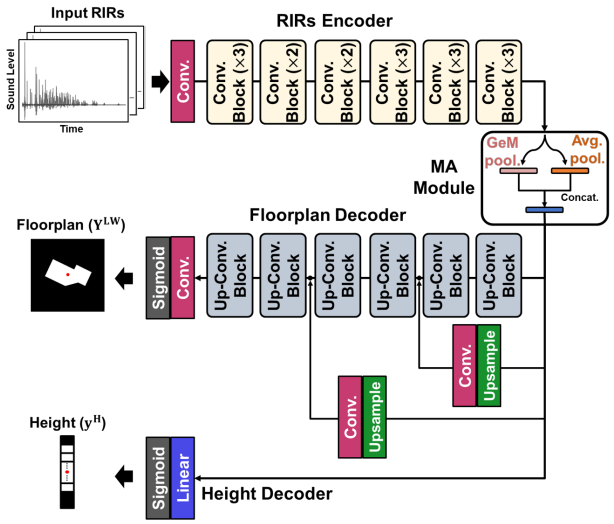

Figure 1. Overview of the proposed model.

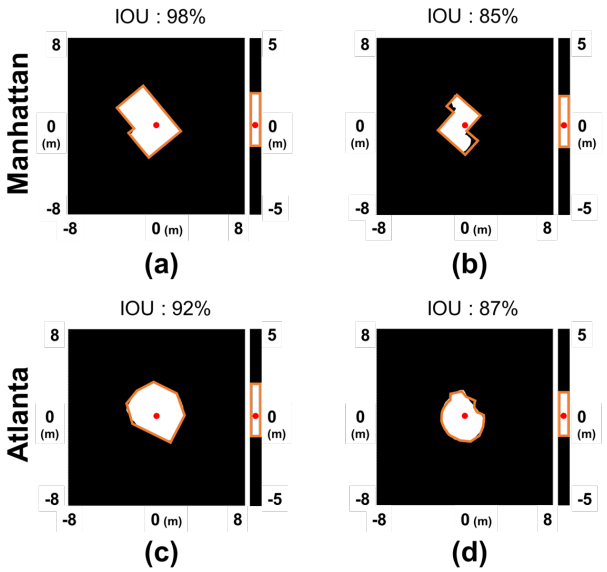

Figure 2. Inferred floorplan and height maps of Manhattan and Altanta layout rooms. The red dot and orange line indicate the device position and the GT room boundary.

height space. Therefore, the acoustic-based RGI task can be approached as a pixel segmentation problem of estimating a 2D floorplan map and a 1D height map containing binary values of 0 or 1 using acquired M-channel RIRs $\mathbf{X} \in \mathbb{R}^{M \times N}$ with temporal length N. The compact acoustic device measuring RIRs is assumed to have a loudspeaker at the center of a circular microphone array with M microphones.

The proposed model employs an encoder–decoder architecture to infer room geometry from RIRs as illustrated in Fig. 1. The encoder processes the multichannel RIRs through the series of convolution blocks, progressively doubling channel dimensions while halving feature dimensions, to convert temporal and inter-channel information into geometry-related features. The MA module compresses the features using multiple pooling operations controlled by the compression parameter ρ [14]. In this study, average pooling ($\rho = 1$) and generalized mean pooling ($\rho = 3$) are utilized to highlight the features activated in global and partially local scales, respectively. These compressed representations are then ensembled to capture both local and global relations in the RIRs. The decoder comprises two specialized components: a floorplan decoder and a height decoder. The floorplan decoder consists of upsampling and convolution blocks with projected skip connections of expanded and reshaped features from the MA module. The height decoder employs a simpler structure with a single linear layer. This is because first-order reflections can always be observed in the assumed geometry with parallel floor and ceiling, and predicting the height map using these first-order reflections is easier than generating the floorplan map. The pixel values of the predicted floorplan and height maps are constrained within the range $[0, 1]$ by applying the Sigmoid activation. During inference, the predicted maps are converted into binary images by hard-thresholding the map using the threshold of 0.5.

The network is optimized by mean squared error (MSE) and dice loss. Dice loss enhances edge details learning by measuring alignment between predicted and ground truth (GT) layouts as

$$L_{\text{dice}}^{\text{LW}} = \frac{1}{I}\sum_{i=1}^{I} 1 - \frac{2(\hat{\mathbf{y}}_i^{\text{LW}})^{\mathsf{T}}\mathbf{y}_i^{\text{LW}}}{\|\hat{\mathbf{y}}_i^{\text{LW}} + \mathbf{y}_i^{\text{LW}}\|_1}, \tag{1}$$

where i is the index of room in the training dataset, and $\mathbf{y}^{\text{LW}}$ is the vectorized form of $\mathbf{Y}^{\text{LW}}$. The total loss function is a joint loss given by $L = L_{\text{MSE}}^{\text{LW}} + 0.3L_{\text{dice}}^{\text{LW}} + L_{\text{MSE}}^{\text{H}}$. Additionally, to address the inherent ambiguity in distinguishing floor and ceiling reflections with circular microphone arrays, permutation invariant training (PIT) [15] is implemented during height map loss calculation.

3. EXPERIMENTAL RESULTS AND ANALYSIS

3.1 Dataset

In this study, an audio device consisting of a circular microphone array with a radius of 5 cm, comprising six omnidirectional microphones and a centrally located loudspeaker was utilized. The device was randomly positioned at heights between $[1, 1.5]$ m from the floor and within 70% of the floorplan of a given room. Multichannel RIRs were simulated using a ray-tracing algorithm provided by Pyroomacoustics [16]. Simulated RIRs have 1024 temporal samples at 8 kHz sampling rate. Background noises (white Gaussian noises) were scaled and added to the RIRs to achieve a signal-to-noise ratio (SNR) randomly selected between $[10, 20]$ dB. Various common acoustic absorbing materials in [16] were randomly assigned to floors, ceilings, and sidewalls for acoustic simulation.

To evaluate the model's capability to infer general and complex room geometries, we utilized five typical room geometries (quadrilateral, pentagonal, hexagonal, L-shaped, and T-shaped) and publicly available room geometry datasets containing diverse Manhattan and Atlanta room layouts [4, 17]. Manhattan layout rooms consist exclusively of walls intersecting at right angles, while Atlanta layout rooms have more general and complex shapes, including curved walls or walls intersecting at oblique angles. To enhance the robustness of the model to possible translation and rotation of the audio device, the center of a floorplan (audio device position) was randomly selected within the area scaled down to 70% of its floorplan and then randomly rotated within the range $[0, 2\pi]$. The final floorplan and height maps were mapped onto a pixel grid of size $b \times b$ ($b = 1024$) and a pixel grid of size h ($h = 512$), respectively, with 2 cm inter-pixel distance.

3.2 Experimental results

The RGI performance of the proposed model is evaluated using intersection over union (IOU). IOU is calculated based on voxel-level overlaps to assess the geometric similarity between the predicted and GT room geometries as

$$\mathrm{IOU} = \frac{1}{I}\sum_{i=1}^{I} \frac{(\hat{\mathbf{y}}_i^{3\mathrm{D}})^{\mathsf{T}}\mathbf{y}_i^{3\mathrm{D}}}{\|\hat{\mathbf{y}}_i^{3\mathrm{D}} + \mathbf{y}_i^{3\mathrm{D}}\|_1 - (\hat{\mathbf{y}}_i^{3\mathrm{D}})^{\mathsf{T}}\mathbf{y}_i^{3\mathrm{D}}}, \quad (2)$$

where $\mathbf{y}^{3\mathrm{D}}$ is the vectorized form of $\mathbf{Y}^{3\mathrm{D}}$.

The RGI performance of the proposed model across five typical room tyeps is presented in Tab. 1. These five room types include both convex (quadrilateral, pentagonal, and hexagonal) and non-convex (L- and T-shaped) shapes, achieving a high IOU of over 90% for all room types. Additionally, the model demonstrates negligible height map estimation errors ($\mathrm{MSE_H}$) regardless of the room type.

Table 1. RGI performance on five typical room types

Evaluation Metric		Convex			Non-convex	
		Quadrilateral	Pentagonal	Hexagonal	L-shaped	T-shaped
IOU (%)	↑	98.5	97.6	97.3	95.7	93.64
$\mathrm{MSE_{LW}}$ ($\times 10^{-3}$)	↓	2.9	3.2	3.5	7.2	10.65
$\mathrm{MSE_H}$ ($\times 10^{-3}$)	↓	0.9	0.8	0.8	0.8	0.9

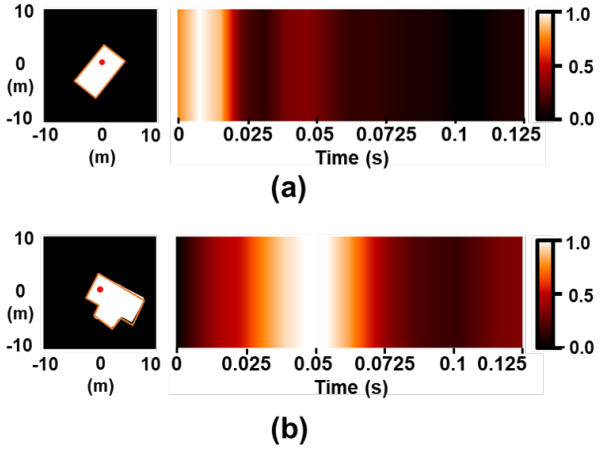

Figure 3. Visualization of temporal activation using Grad-CAM for (a) quadrilateral room and (b) T-shaped room. The red dot and orange line denote the device position and the GT room boundary.

Fig. 2 demonstrates inference results of the floorplan and height maps of Manhattan and Atlanta layout rooms. These examples indicate the proposed model captures the overall layout of complex room geometries with small errors around corners. Moreover, the inference result of Fig. 2(d) shows that the model can accurately infer drastically curved geometry.

Exploiting high-order reflections is essential for inferring NLOS walls. To verify that the proposed model utilizes high-order reflections, we visualize temporal activation maps of RIRs using gradient-weighted class activation mapping (Grad-CAM) [18]. For simple quadrilateral rooms (Fig. 3(a)), the model highlights early temporal regions dominated by low-order reflections. In contrast, for T-shaped rooms with NLOS walls (Fig. 3(b)), stronger activations are observed in later temporal regions around

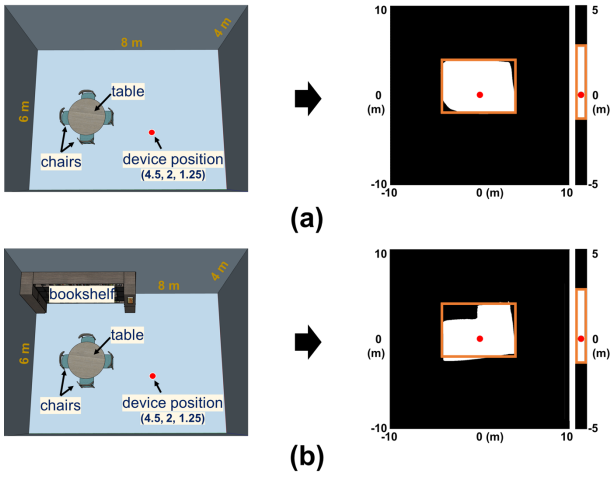

Figure 4. The inference results when the indoor objects are present. The left panels display a top view of 3D models, and the right panels show the inferred floorplan and height maps. The red dot and orange line illustrate the device position and the GT room boundary.

0.05 s corresponding to higher-order reflections. These results suggest that the proposed model effectively leverages higher-order reflections to estimate complex room geometries.

The RIRs used for training are simulated in empty rooms. However, indoor objects such as furniture can influence the RIRs. Figure 4 shows the 3D models of a quadrilateral room (left) with dimensions $(8, 6, 4)$ m containing indoor objects, and their inference results (right). These results indicate that objects shorter than the audio device position (chairs and table) have minimal impact on the inferred floorplan, while taller object (bookshelf) is identified as wall. These results suggest that the proposed model can infer room geometry even when the acquired RIRs are affected by indoor objects.

4. CONCLUSION

This study presents a DNN-based RGI model using acoustic echoes. Unlike room parameter estimation approaches, the proposed pixel segmentation approach has demonstrated its ability to handle a wide range of room shapes. Moreover, the experimental results visually demonstrate that the proposed model effectively exploits high-order reflections to accurately reconstruct room geometries even in cases where some walls are not directly visible from the audio device position.

5. ACKNOWLEDGMENTS

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Ministry of Science and ICT of Korea government (MSIT) (No. RS-2024-00337945), the BK21 FOUR program through the NRF grant funded by the Ministry of Education of Korea government (MOE).

6. REFERENCES

[1] S. Werner, F. Klein, T. Mayenfels, and K. Brandenburg, "A summary on acoustic room divergence and its effect on externalization of auditory events," in *Proc. IEEE Int. Conf. Quality Multimedia Experience (QoMEX)*, pp. 1–6, IEEE, 2016.

[2] I. Dokmanić, R. Scheibler, and M. Vetterli, "Raking the cocktail party," *IEEE J. Sel. Top. Signal Process.*, vol. 9, no. 5, pp. 825–836, 2015.

[3] C. Sun, C.-W. Hsiao, M. Sun, and H.-T. Chen, "Horizonnet: Learning room layout with 1d representation and pano stretch data augmentation," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, (Long Beach, CA, USA), pp. 1047–1056, 2019.

[4] G. Pintore, M. Agus, and E. Gobbetti, "Atlantanet: inferring the 3d indoor layout from a single 360° image beyond the manhattan world assumption," in *Proc. Eur. Conf. Comput. Vis.*, (Glasgow, UK), pp. 432–448, Springer, 2020.

[5] F. Antonacci, J. Filos, M. R. Thomas, E. A. Habets, A. Sarti, P. A. Naylor, and S. Tubaro, "Inference of room geometry from acoustic impulse responses," *IEEE Trans. Audio, Speech, Language Process.*, vol. 20, no. 10, pp. 2683–2695, 2012.

[6] I. Dokmanić, R. Parhizkar, A. Walther, Y. M. Lu, and M. Vetterli, "Acoustic echoes reveal room shape," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 110, no. 30, pp. 12186–12191, 2013.

[7] Y. El Baba, A. Walther, and E. A. Habets, "3d room geometry inference based on room impulse response stacks," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 26, no. 5, pp. 857–872, 2017.

[8] S. Park and J.-W. Choi, "Iterative echo labeling algorithm with convex hull expansion for room geometry estimation," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 29, pp. 1463–1478, 2021.

[9] M. Lovedee-Turner and D. Murphy, "Three-dimensional reflector localisation and room geometry estimation using a spherical microphone array," *J. Acoust. Soc. Am.*, vol. 146, no. 5, pp. 3339–3352, 2019.

[10] W. Yu and W. B. Kleijn, "Room acoustical parameter estimation from room impulse responses using deep neural networks," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 29, pp. 436–447, 2020.

[11] C. Tuna, A. Akat, H. N. Bicer, A. Walther, and E. A. Habets, "Data-driven 3d room geometry inference with a linear loudspeaker array and a single microphone," in *Proc. Eur. Acoust. Assoc. (Forum Acusticum 2023)*, (Torino, Italy), 2023.

[12] I. Yeon and J.-W. Choi, "Rgi-net: 3d room geometry inference from room impulse responses with hidden first-order reflections," in *Proc. Int. Workshop Acoust. Signal Enhance. (IWAENC)*, (Aalborg, Denmark), pp. 439–443, IEEE, 2024.

[13] I. Yeon, I. Jeong, S. Lee, and J.-W. Choi, "Echoscan: Scanning complex room geometries via acoustic echoes," *IEEE/ACM Trans. Audio, Speech, Language Process.*, 2024.

[14] F. Radenović, G. Tolias, and O. Chum, "Fine-tuning cnn image retrieval with no human annotation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 7, pp. 1655–1668, 2018.

[15] D. Yu, M. Kolbæk, Z.-H. Tan, and J. Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, (New Orleans, LA, USA), pp. 241–245, 2017.

[16] R. Scheibler, E. Bezzam, and I. Dokmanić, "Pyroomacoustics: A python package for audio room simulation and array processing algorithms," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, (Calgary, AB, Canada), pp. 351–355, IEEE, 2018.

[17] C. Zou, J.-W. Su, C.-H. Peng, A. Colburn, Q. Shan, P. Wonka, H.-K. Chu, and D. Hoiem, "Manhattan room layout reconstruction from a single 360° image: A comparative study of state-of-the-art methods," *Int. J. Comput. Vis.*, vol. 129, no. 5, pp. 1410–1431, 2021.

[18] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis.*, (Venice, Italy), pp. 618–626, 2017.