Manchester Metropolitan University

e-space

Manchester Metropolitan University's Research Repository

Please cite the Published Version

Colombo, Mattia, Dolhasz, Alan, Hockman, Jason and Harvey, Carlo (2022) Acoustic rendering based on geometry reduction and acoustic material classification. In: 2022 IEEE Conference on Games (CoG), 21 August 2022 - 24 August 2022, Beijing, China.

DOI: https://doi.org/10.1109/cog51982.2022.9893613

Publisher: IEEE

Version: Accepted Version

Downloaded from: https://e-space.mmu.ac.uk/633021/

Usage rights: © In Copyright

Additional Information: © 2022 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works

Enquiries:

If you have questions about this document, contact openresearch@mmu.ac.uk. Please include the URL of the record in e-space. If you believe that your, or a third party's rights have been compromised through this document please see our Take Down policy (available from https://www.mmu.ac.uk/library/using-the-library/policies-and-guidelines)

Acoustic Rendering Based on Geometry Reduction and Acoustic Material Classification

***Abstract*—We present work in progress on a pipeline for audio rendering integrating vision-based systems for acoustic material classification. With a marching cubes algorithm, the pipeline estimates a cuboid acoustic volume encapsulating the listener, a sound source, and the surrounding environment. A variable-resolution binary field, samples and simplifies the input scene, and captures the appearance of surfaces to produce a set of image patches. A classifier infers acoustic materials, expressed as frequency-dependent acoustic absorption coefficients, from image patches. The estimated volume and aggregated acoustic materials provide input to the Image Source Model that models reverberation by generating Room Impulse Responses (RIRs). We conduct preliminary tests by applying our pipeline on a set of indoor and outdoor scenes, producing RIRs with inferred acoustic materials, comparing them against RIRs with manually assigned acoustic materials, extracting and evaluating objective metrics, such as reverberation or clarity. Through a learned metric on subjective responses, we compare perceptual aspects of automatically-generated RIRs against manually tagged. Objective and subjective analysis suggests that the pipeline can automate the acoustic material classification process by producing RIRs indistinguishable from manually-tagged counterparts.**

***Index Terms*—acoustic rendering, material recognition, interactive applications**

I. Introduction

Thanks to the increasing popularity of sound rendering, it is now possible to produce realistic acoustic models, where sound propagates in virtual environments, interacting with boundaries and objects, finally arriving at the listener's ears [1], [2]. Virtual environments and interactive applications benefit from simulated sound propagation, as it enables rich multimodal interactions, contributes to evoking a sense of presence and immersion [3], [4], and improves task performance [5]. Acoustic rendering aims at reconstructing virtual acoustic space, replicating characteristics of a real environment, by generating artificial RIRs across positions of sound sources and listeners, which audio engines can use to propagate sound from sources to the user's ears.

Input to sound propagation methods consists of scene geometry, acoustic properties of surfaces and boundaries, and the position of sources and listeners in space [6], [7]. They are determinants of the resulting perceived auralisation quality and represent a crucial challenge of audio rendering, and typically require human intervention in the scene authoring process, especially for realistic acoustic simulations. Determining acoustic characteristics of surfaces requires the work of expert acousticians, an obstacle to real-time rendering, which has gained increasing attention in the field of sound rendering. Recent work has approached this problem by adopting cutting edge machine learning techniques [8]–[11].

Our goal is to explore scalable deep learning techniques for interactive and plausible sound rendering, combining fast methods of sound propagation with automatic input generation to produce plausible acoustic simulations in virtual environments. Given a complex scene, we aim to capture the appearance of objects and surfaces to map their visual features to acoustic characteristics, drawing from Schissler *et al.*'s work. Our vision advances this paradigm by optimising the visual information sampled from scenes based on the perceived quality of the resulting acoustic simulation. Here, we present a prototype acoustic rendering pipeline, targeting platforms with dynamic geometry by integrating automatic acoustic material recognition within the process of reconstructing the acoustic environment surrounding a listener and a sound source, allowing for automatic sound rendering.

We report on a system that captures of the environment's appearance, using material classification network to predict acoustic parameters of the scene from its appearance, expressed as image patches. A surface reconstruction algorithm samples the visuals of the environment to generate the image patches. The current state of the system has acoustic materials aggregating to mean materials assigned to the cuboid volume of the Image Source Model (ISM). The ISM is a GA method that computes acoustic reflections, modelling sound propagation from a sound source to a listener. The next step of this system is to combine the reflection computation with the surface reconstruction process. Hence, the system would calculate reflections based on the surface intersected, using acoustic characteristics extracted from the visuals of the surface. This approach would also overcome the problem of the cuboid acoustic volume.

II. Related Work

In acoustics, it is common to capture an environment adopting measurement techniques such as the sine sweep, usually consisting of reproducing a logarithm sine chirp or a short burst, i.e. a gunshot, emulating a Dirac delta function to excite frequencies in the audible spectrum and recording how the environment influenced the propagated sound at the listener position [12]. Such measurements can determine a Room Impulse Response (RIR), a series of reflection paths over time, recreating the acoustic space for a given source-listener position pair. Wave-based acoustic simulations achieve the highest degrees of realism in generating acoustic fields as they compute sound propagation via simulations of high-dimensional pressure fields [13] or solving the wave equation with Finite-Difference Time-Domain schemes [14]. Their in-

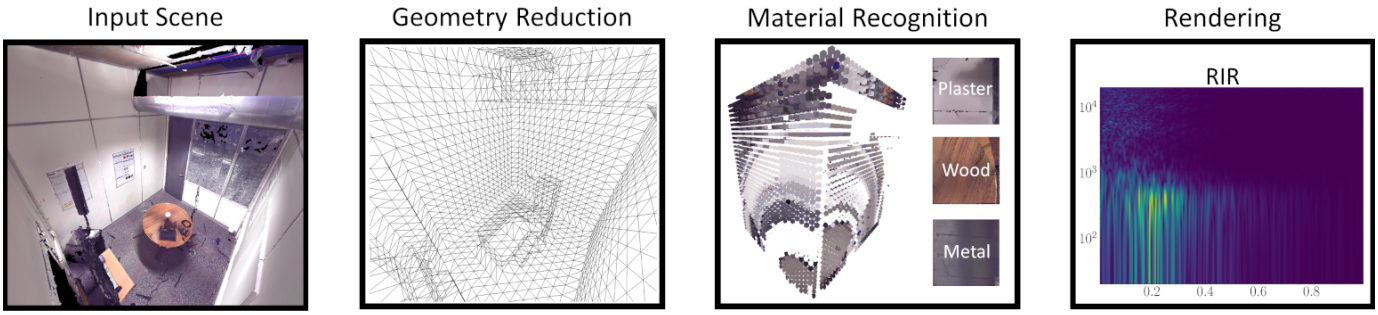


Fig. 1. Breakdown of the proposed acoustic renderer. From left to right: an input scene is obtained from the input platform and represented as textured meshes; the scene geometry is reduced via a geometry reduction process, determining the acoustic volume; material patches are generated from textures, depending on size and position of each cell in the acoustic volume; the acoustic renderer generates RIRs that can be convolved to audio emitted from a sound source in the scene.

herently complex nature require solving the wave equation to produce acoustic simulations for a given scene, and despite recent GPU-based solvers optimising complexity by orders of magnitude [15]. Their computational requirements are often impractical for real-time applications due to the nature of the wave equation, resulting in numerical complexity increases with frequency. On the other end of the spectrum, Geometrical Acoustics (GA) provide methods for fast approximations of acoustic space; they have gained popularity among extended reality platforms due to their highly parallelisable implementations [16].

A. Acoustic Materials

Sound rendering methods simulate model phenomena of sound propagation, such as reflection or diffraction based on physical properties of scene geometry, which are determinant factors of a sound field. Among these, the T_{60} acoustic metric indicates the time required for acoustic energy to decay by $60dB$, and it is commonly used in measuring sound fields as a function of volume and surface with relative absorption. The frequency-dependent T_{60}, in addition, explains reverberation over a frequency spectrum, usually across six frequencies in the equivalent rectangular bandwidth scale to approximate the human hearing range [17]. In room acoustics, the T_{60} and other metrics such as C_{50} or the D_{50}, which indicate the clarity and definition index of a sound field [18], are dependent on architectural features and their materials, among other factors. In acoustic rendering, materials are often defined as features attributed to virtual geometry composing a complex scene to model interaction between geometry and propagating sound waves. Acoustic absorption, for instance, can determine how much energy virtual geometry reflects given a colliding incidental wave.

B. Acoustic Rendering

Schissler and Manocha [19] introduced an acoustic rendering system based on ray-tracing, adapting to large complex scenes. Among their contributions is overcoming the problem of handling many sound sources in large-scale environments by clustering them based on the distance from the listener. Based on an octree representation of space, with respect to the listener position, their clustering aggregates increasing numbers of sources as their distance from the listener increases. Their approach highlights the need for dissecting the acoustic space for efficient selective rendering, resulting in rendering of fine perceptual details within the listener's close proximity and coarse approximations otherwise.

Vorländer's work [20] pioneered the combination of the image-source model with ray tracing to prototype sound propagation for interactive applications, and one of the first forms of geometry reduction for acoustic rendering appeared in Siltanen *et al.*'s work [21]. They considered an isosurface extraction algorithm, marching cubes, to simplify complex, unstructured models, producing acoustic simulations via geometrical acoustics within minutes. To the best of our knowledge, this approach has not been developed further to consider recent advances in scene understanding applied to virtual reconstructions of environments to recognise acoustic characteristics of space, such as acoustic materials [22], which can be captured from complex scenes.

Schissler *et al.* [23] recently introduced a novel acoustic renderer that reconstruct real-world scenes, using 3D reconstructions from a set of camera viewpoints. By leveraging image segmentation, they infer the semantics of scene objects to produce and optimise acoustic materials assigned to the scene. The renderer generates impulse responses that are processed with delay interpolation, head-related transfer functions and panning to adapt to human listeners. The above-mentioned works demonstrate how a network can learn complex mappings between cross-modal stimuli and acoustic characteristics of space, producing input to sound propagation methods.

C. Material Recognition

In-the-wild datasets, such as OpenSurfaces [24], and Matterport3D [25], capture material appearances in surfaces with sub-optimal lighting conditions, specular reflections and errors introduced by optic sensors, which may result in noisy image representations of materials. Classification of image patches through densely connected networks can achieve reasonable accuracy when trained on extracted features, with the condition that the data has a comprehensive representation of materials applicable to target scenes of the acoustic rendering pipeline,

as shown in recent work by Colombo *et al.* [26]. Based on captured real environments, the authors automated the mapping between the visual representation of materials, expressed as meshes whose textures are decomposed into superpixels to infer their acoustic characteristics via a learned classifier.

Gaur and Manjunath [27] present a novel clustering network that would provide an alternative to classification paradigms in acoustic material tagging tasks. Their novel technique subdivides an input image into superpixels, image segments whose pixels belong to similar texture patterns, by extracting pixel embeddings using unsupervised autoencoders. Their work allows the clustering of features across entire datasets. Acoustic material tagging would benefit from this technique by removing the need for semantic labels, potentially allowing interpolation between acoustic materials across clusters of features extracted from a dataset of surface appearances.

III. Method

The proposed framework employs a GA method, the Image-Source Model [28], [29], to compute reflection paths in a given virtual scene reconstruction, as illustrated in Figure 1. The *geometry reduction* component of this pipeline decomposes the virtual scene into a simpler cuboid volume. Vision-based material recognition techniques are used to understand acoustic characteristics of surfaces in the input scene; these acoustic characteristics are then mapped onto the appropriate surfaces of the acoustic volume that represents the given virtual scene. Finally, the ISM is supplied with spatial information on emitters and listeners to produce RIRs from the acoustic volume.

A. Geometry Reduction

Our *geometry reduction* system generates a binary field, sampling from the input geometry at a given number of points. The scalar field determines the acoustic volume via a marching cubes algorithm, where each cell captures the appearance of surfaces from textures associated with meshes of the input geometry. A cuboid volume encapsulating the reconstructed surface determines the dimensions of the acoustic space simulated with the ISM. Using vision-based techniques to analyse surface appearance, we infer an acoustic material for each image patch, effectively mapping acoustic materials to the surface of the acoustic volume.

1) Volume Estimation: Input geometry is decomposed into a binary field expressed as $\mathbf{B} = (b_{x,y,z}) \in \{0,1\}^P$, where P is the number of points sampled across the three dimensions. The definition of $\mathbf{B}$ is relevant to the quality of the resulting acoustic simulation and determines the accuracy of the acoustic material recognition, discussed in the following section. Given the set of input meshes from the complex scenes, we define an Axis-Aligned Bounding Box (AABB) encapsulating all vertices of a given scene object, normalising their coordinates to the $[-1,+1]$ range. Values in $\mathbf{B}$ equate to 1 whenever a cell is within the coordinates of any AABB defined, and to 0 otherwise. We will refer to AABBs as the collection of all bounding boxes associated with scene objects throughout the rest of the paper. A multi-threaded implementation of the marching cubes algorithm [30], [31] shapes the binary field into a volume, representing the acoustic environment for reflection path computation. In this process, the use of the AABB to probe the input scene objects introduces error in cases with concave geometry, which negatively correlates to the resolution of $\mathbf{B}$ as collision checks with AABBs increase with the number of cells. A cuboid encapsulating the reconstructed surface from the marching cubes determines the ISM volume for the reflection computation stage.

B. Acoustic Material Tagging

We introduce a camera-based system that generates a set of image patches. These patches are generated from orthonormal projections of candidate marched cubes intersecting the reconstructed surface. The image patches feed a neural network to extract features that are then classified into an acoustic material. Hence, each image patch corresponding to an acoustic material defines the acoustic properties of portions of the complex scene. Based on the space subdivision computed at the *geometry reduction* stage, our system samples the appearance of surfaces to recognise their semantics, which maps to physical characteristics describing their behaviour when interacting with reflecting sound waves.

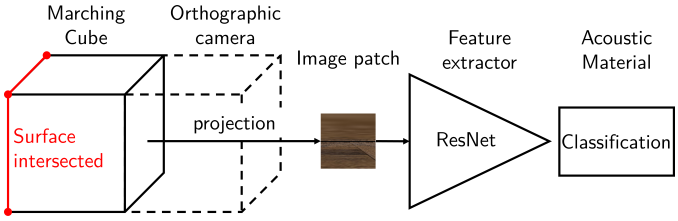


Fig. 2. Image patch generation: surfaces intersected by marching cubes are projected onto image patches with a camera by rasterising vertices via orthographic projections of the delimited portion of surface. The resulting image patch is then fed to a neural network to extract features for semantic classification. Through a one-to-many mapping, semantics attributed to a patch, via embeddings classification, map to acoustic parameters.

C. Camera Projection

Image patches are generated whenever a marching cube intersects a surface by an entire side, i.e. four consecutive corners of each marching cube. An orthographic camera, positioned at the centre of the neighbouring cube, rasterises sub-surfaces delimited by the cube. The camera is positioned in the opposite side of the four intersected corners, facing the surface that is intersected, see Figure 2. With OpenGL rasterisation, we transform vertices of surfaces to the world space, with clipping based on the volume of the marching cube. Sampling image data from textures, the rasterisation stage produces an image patch representing the projection of the portion of surface inscribed by a marching cube. Marching cubes intersecting in four consecutive corners ensure that the orthographic camera is perpendicular to the surface.

D. Acoustic Material Classifier

Image patches projected from marching cubes provide input to a neural network, a ResNet50 [32] backbone that operates

as a feature extractor, whose output is forwarded to a densely connected layer. This last layer classifies predicted embeddings into acoustic material categories. The network was trained on the OpenSurfaces dataset [24], learning mappings between visual appearances of surfaces from 34 material classes and semantic labels, as described in [26]. The network, pre-trained on ImageNet [33], learns on 32x32 pixel resolution image patches that we extract from appearances sampled from OpenSurfaces, assembling a dataset of about 13M images, split into 9M and 4M train and evaluation sets, respectively. The model converges in 45 epochs, achieves validation accuracy of about 0.83, and, as shown in recent experiments [26], it can replace humans in the process of acoustic material tagging in enclosed scenes of real-world space. Via a one-to-many mapping, we associate each of the 34 categories from the OpenSurfaces dataset to acoustic materials. Given 11 acoustic materials, subdividing into two levels of mass density, the visual labels are mapped to 22 acoustic materials. Model architecture and weights are available[1].

E. Frequency-dependent geometrical acoustics

The ISM estimates attenuation functions for a given listener position, with respect to a source in space, drawing from Habet's implementation [34] to estimate h energy functions, based on the spatial $[x, y, z]$ coordinates of a sound source, $\mathbf{s}$, a listener, $\mathbf{l}$, at each time step t as $h(\mathbf{l}, \mathbf{s}, t) =$

$$\sum_{\mathbf{p}\in\mathcal{P}} \sum_{\mathbf{m}\in\mathcal{M}} \beta_{-x}^{|m_x-q|} \beta_{+x}^{|m_x|} \beta_{-y}^{|m_y-j|} \beta_{+y}^{|m_y|} \beta_{-z}^{|m_z-k|} \beta_{+z}^{|m_z|} \frac{\delta(t-\tau)}{4\pi d} \tag{1}$$

Here, $\mathcal{M} = \{(m_x, m_y, m_z) : -N \leq m_x, m_y, m_z \leq +N\}$ determines the number of points per dimensions based on the size of the input acoustic volume and the timestep, dependent on the desired sampling frequency, on the length of the output RIR, and the order of reflections N. $\mathcal{P} = \{(q, j, k) : q, j, k \in 0, 1\}$ determines possible combinations of image sources mirrored in the three dimensions from every boundary in the acoustic volume to consider higher-order reflections, which are computed by $\delta(t - \tau)$, where $\tau = \frac{||\mathbf{R_p}+\mathbf{R_m}||}{c}$ indicates the reflection time delay by dividing the measured distance between mirrored image positions $\mathbf{R_p} + \mathbf{R_m}$ and the listener by the speed of sound c, d is the distance term and is calculated as $\sqrt{(\mathbf{R_m} + \mathbf{R_p})^2}$. With combinations in $\mathcal{P}$, image positions are determined by $\mathbf{R_p} = [(1 - 2q)\mathbf{s_x} - \mathbf{l_x}, (1 - 2j)\mathbf{s_y} - \mathbf{l_y}, (1 - 2k)\mathbf{s_z} - \mathbf{l_z}]$ and $\mathbf{R_m} = [2m_x l_x, 2m_y l_y, 2m_z l_z]$. The ISM computes multiple h functions across frequency-dependent reflection coefficients, increasing the accuracy of simulated reflections from boundaries of the acoustic volume. In Equation 1, β reflection coefficients determine the energy attenuation of a computed reflection, specifying a single reflection coefficient mapping to each side of the acoustic volume; namely, $-x$ to $+z$ (left, right, top, bottom, front and back side). As these coefficients imply that materials apply a constant attenuation over the frequency spectrum, we redefine them as $\beta_{-x,f} \ \ldots \ \beta_{+z,f} \ \forall f \in \mathcal{F} : \{125, 250, 500, 1000, 2000, 4000\}Hz$, where f indicates the frequency bin mapping to a reflection coefficient, adding a further dimension to Equation 1, which can be defined as $h(\mathbf{l}, \mathbf{s}, f, t)$ equating to:

$$\sum_{\mathbf{p}\in\mathcal{P}} \sum_{\mathbf{m}\in\mathcal{M}} \sum_{f\in\mathcal{F}} \cdot$$
$$\beta_{-x,f}^{|m_x-q|} \beta_{+x,f}^{|m_x|} \beta_{-y,f}^{|m_y-j|} \beta_{+y,f}^{|m_y|} \beta_{-z,f}^{|m_z-k|} \beta_{+z,f}^{|m_z|} \frac{\delta(t-\tau)}{4\pi d} \tag{2}$$

Frequency-dependent reflection coefficients enable the mapping between boundaries in the environment and acoustic materials. Equation 2 produces separate h attenuation functions for frequency bins in $\mathcal{F}$, associated to the common six octave bands defined by the acoustic materials to cover the equivalent rectangular bandwidth-number scale [16], [17].

F. Audio Rendering

We process h functions generated by convolving them finite impulse response filter. We design six filters based on frequency-dependent absorption coefficients to obtain a new set of h functions that contribute to a specific band of the equivalent rectangular bandwidth scale. Phase-invariant low-pass filters based on Smith *et al*'s [35] designs are combined with their corresponding frequency-inverted counterparts that we chain to produce band-pass filters. Filters are complementary to each other in the frequency domain ($20Hz$ to $20kHz$), summing to a flat magnitude response. Processed h functions are then summed into a resulting RIR, which we convolve to anechoic audio to propagate audio in the simulated acoustic environment.

G. Acoustic Volume Absorption

We adapt our ISM to non-enclosed or partially enclosed space by determining acoustic materials associated with the six sides of the cuboid acoustic volume: when no image patches are assigned to a given side, its respective reflections are ignored, i.e maximum attenuation. Otherwise we let $\mathbf{M}_{-x \ \ldots \ +z} = \{\alpha_{0,f}, \ \ldots \ , \alpha_{n,f}\}$ be the vectors of acoustic materials corresponding to marching cubes, describing frequency-dependent α acoustic absorption. With marching cubes intersecting surfaces associated with a side of the volume at n points, acoustic materials generated by the *material recognition* stage substitute elements of vector $\mathbf{M}$, while the remaining elements default to air absorption [36]. Acoustic materials contribute to a final set of acoustic absorption coefficients $\alpha_{-x,f} \ \ldots \ \alpha_{+z,f}$ equivalent, as $\beta = \sqrt{1 - \alpha}$ [29], to the reflection coefficients considered in Equation 2. The contribution of each acoustic material is weighted based on the number of image patches found for each side and the total number of possible image patches P^3 that can be associated to each side, dependent on the resolution of the *geometry reduction* stage. Hence, the weighted average determining

[1]https://anonymous.4open.science/r/marching-cubes-acoustic-rendering-4BE4/README.md

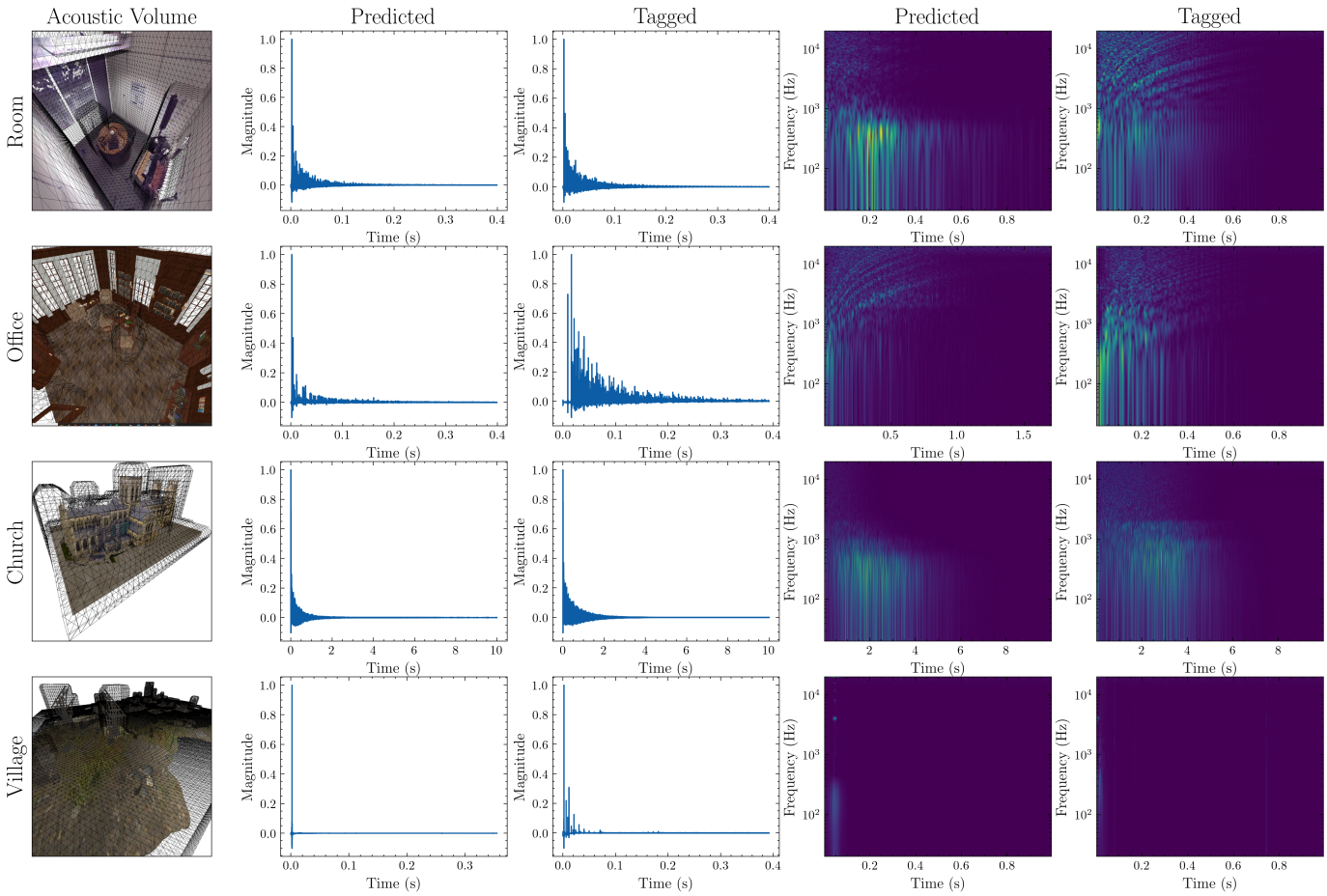


Fig. 3. Comparison between room impulse responses, generated using the proposed framework. *Predicted* RIRs are produced using materials inferred by the automatic acoustic material classification, while the *tagged* counterparts have manually tagged acoustic materials. Rows show scenes in ascending order of volume; Columns from left to right show: a render of the scene with an overlapped polygonised acoustic volume resulting from the marching cubes algorithm; a time-domain visualisation of the impulse response using *predicted* materials, followed by the counterpart with *tagged* materials; finally, the last two columns show spectrograms of the two. RIR pairs are generated maintaining the same positions of source and listener.

acoustic absorption for each side of the acoustic volume can be defined as:

$$\alpha_{-x,f} \;\; \dots \;\; \alpha_{+z,f} = \frac{\sum_{i=1}^{P^3} w_i \; \mathbf{M}_{-x,i} \;\; \dots \;\; \mathbf{M}_{+z,i}}{\sum_{i=1}^{P^3} w_i}, \tag{3}$$

where w indicates the weights vector defining acoustic material contribution:

$$w_i = \begin{cases} \frac{n}{P^3} & \text{if } i \leq n \\ \frac{(P^3 - n)}{P^3} & \text{otherwise} \end{cases} \tag{4}$$

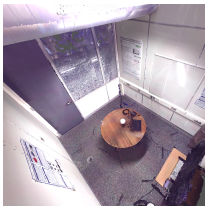
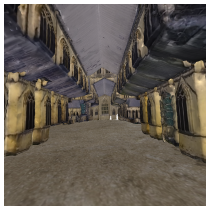
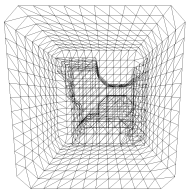
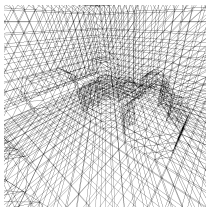
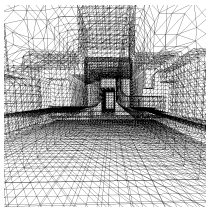
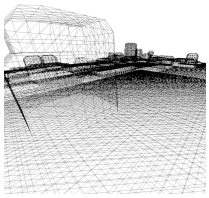
Hence, the contribution of acoustic materials to the ISM volume depends on the surface intersected by marching cubes.

H. Perceptual Evaluation

We conduct preliminary tests by deploying the proposed pipeline on a set of scenes, where we place a listener and sound source. The pipeline constructs a cuboid acoustic volume encapsulating the source-listener pair, produces image patches with captured surfaces and infers acoustic materials to generate acoustic materials associated with the volume. For each scene, we generate two RIRs: one with automatically-generated acoustic materials and one with manually-tagged acoustic materials. We compare RIRs by utilising a pre-trained network for subjective comparison between auralisations produced via convolution of RIRs to audio from a database of anechoic recordings of sound events from the TUT Sound Event database [37]. The deep audio perceptual audio similarity metric, CDPAM [38], trained on a dataset of human judgements, expresses distances between two signals and a reference. Hence, the metric can be used to measure Just Noticeable Differences (JND) as a distance $D(x_{per}, x_{ref})$ between a perturbed signal x_{per}, and a reference x_{ref}. Reverberation or equalisation are among the perturbation factors affecting the measured distance. Perceptual distance scores greater than 1 indicate that a human would distinguish them as distinct. The objective of the evaluation is to evaluate whether sound propagated using the rendering pipeline with inferred acoustic materials is perceptually indistinguishable from sound propagated with manually-tagged materials. We convolve to a collection S of 2700 audio samples to then determine

TABLE I

FEATURES EXTRACTED FROM ROOM IMPULSE RESPONSE PAIRS GENERATED USING THE PROPOSED SYSTEM AND METRICS OF CORRESPONDING INPUT ENVIRONMENTS. EACH PAIR HAS A *predicted* AND *tagged* RIR, REFERRING TO ACOUSTIC MATERIALS BEING INFERRED WITH ACOUSTIC MATERIAL CLASSIFICATION OR TAGGED MANUALLY. t REFERS TO THE TIME TAKEN TO COMPUTE REFLECTIONS, AND P INDICATES THE NUMBER OF SAMPLING POINTS PER DIMENSION.

	Room	Office	Church	Village
Scene				
Acoustic Volume				
Volume (m^3)	3.43×10	8.64×10^3	4.83×10^4	3.45×10^8
Triangles	15.4M	0.973M	0.49M	9.6M
P	16^3	32^3	64^3	128^3
Order	3	4	5	1
$t(s)$	2.824	3.53	2.791	1.25
predicted $T_{60}(s)$	0.997	1.705	5.982	0.103
tagged $T_{60}(s)$	0.331	0.734	9.6	0.124
T_{60} error	0.666	0.971	3.618	0.021
predicted C_{50}	1.278	-2.613	-12.341	-5.724
tagged C_{50}	0.709	-1.193	-13.706	-3.036
C_{50} error	0.569	1.42	1.365	2.688
predicted D_{50}	0.499	0.384	0.048	0.275
tagged D_{50}	0.509	0.619	0.035	0.36
D_{50} error	0.01	0.235	0.013	0.085

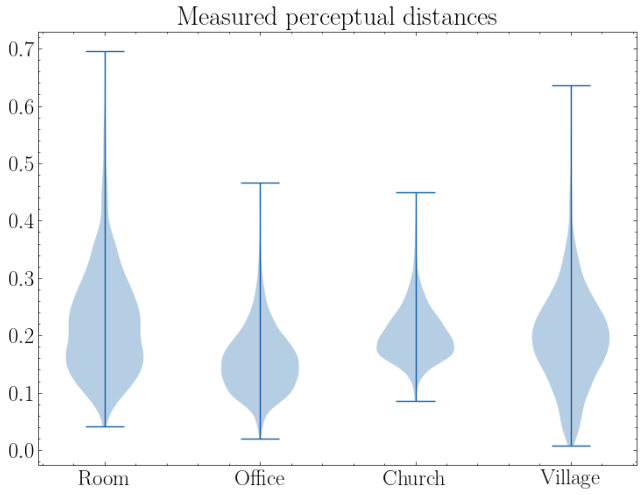


Fig. 4. Distributions of perceptual distances between pairwise comparisons of audio recordings convolved to generated RIR pairs. A collection of everyday sounds is propagated for each pair, producing convolution pairs. By employing a learned metric, the perceptual distance between RIR with *predicted* and *tagged* materials is measured. The violin plot reports that all distances fall below one just noticeable difference, indicating that a human would be unlikely to distinguish between the two convolutions.

perceptual distances $D(x_{predicted,i}, x_{tagged,i})\ \forall i \in S$ between convolutions using predicted materials and manually tagged materials.

IV. RESULTS

We test the system by deploying the proposed pipeline on a set of scenes, providing coverage of a wide range of practical use cases: “Room”, “Office”, “Church” and “Village”. These have increasing volume as reported on Table I. “Room” is a real conference room captured using a LiDAR scanner, FARO Focus^{3D} X300. The remaining scenes are virtual environments that are common in computer games development.

Using our framework, RIRs are generated across the four scenes, applying all stages of our pipeline as offline procedures, comparing acoustic simulations with inferred acoustic materials to simulations with manually-tagged materials. These acoustic simulations assume omnidirectional sources and receivers as no human listeners are involved in the comparison. Hence, simulations disregard directivity patterns or head-related transfer functions. Source-receiver position pairs are consistent across RIRs pairs. We extract objective metrics associated with room acoustic parameters to compare

the output of each acoustic simulation. The classifier takes an average of $0.03s$ to infer the acoustic material from an image.

We evaluate our sound rendering pipeline comparing acoustic energy decay across computed RIRs, as it expresses reflection paths computed on scene geometry for a given source-listener position pair. We extract metrics from impulse responses following De Lima *et al.*'s feature analysis definitions [18]. By fitting energy decay curves, we determine the T_{60} reverberation metric, the C_{50} clarity index, and the D_{50} definition index. C_{50} and D_{50} indices are dependent upon the ratio between the power of early and late reflections. See Table I for estimated reverberation, clarity and definition scores across scenes. Figure 4 show distributions of perceptual distances from RIRs generated using automatically tagged materials to RIRs generated with manually assigned materials.

V. Discussion

A. Overview

Given that procedures of isosurface extraction and computation of frequency-dependent impulse responses run on CPU programs, the timings recorded across the four scenes with increasing spatial resolution suggest that the pipeline would be practical for interactive platforms. Especially when considering dynamic and incremented geometry typical of extended reality systems. Furthermore, GPU implementations of the ISM efficiently distribute sources across parallel workers, allowing for real-time RIR generation [39]. The technique illustrates promise in the domain of streaming geometry, where virtual environments are constructed via spatial mapping services for visualisation on XR displays. There is a necessity for understanding the acoustic information to be associated with this streamed mesh data. The nature of this approach facilitates mesh ingestion and updates to the binary field, subsequent subregions of the marching cube volume can be iteratively updated, and the resultant RIR to take stock of updated geometry can thus be generated. Despite overcoming the limitations of ISMs in propagating sound in non-enclosed environments, phenomena such as occlusion of sound sources and arbitrary shapes of the environments are not considered. Occlusion and visibility of sound sources can be solved by combining the ISM with ray tracing, allowing for checking source visibility introducing limited overhead thanks to their GPU implementations [40].

B. Acoustic Material Approximations

The complexity of the pipeline depends upon the nature of the complex scenes and the number of points in which surfaces are sampled. Considering the overhead introduced by the acoustic material classifier, the complexity of the pipeline scales linearly with the number of surface intersections in the environment. Hence, the worst complexity occurs when each marching cube intersects scene geometry. In addition, despite the classifier's reasonable accuracy on test data, no ablation studies have been conducted to reduce the architecture to a minimum topology and further optimise complexity. Acoustic materials associated with each side of the calculated volume contribute to a mean, causing the ISM to approximate the computation of specular reflections, neglecting characteristics of surfaces, such as position or orientation. This approximation can be overcome by eliminating the process of averaging acoustic materials and reformulating the computation of image sources defined in Equation 2 to refer to individual acoustic materials mapped to the acoustic volume. However, the reformulation would affect the timestep of computed h functions, constraining it to the geometry reduction resolution, resulting in an arbitrary scale that should be interpolated to the time scale. The benefits of removing mean acoustic material would include the ability to simulate arbitrary shapes of surfaces by having acoustic materials mapped to marching cubes. This process would need to consider Nyquist sampling theory to determine the appropriate cube sizes to simulate accurate acoustics whilst maintaining specular plausibility in the frequencies simulated. Pelzer and Vorländer's work [41], in addition, suggests that the resolution of the geometry reduction process can be set according to perceptual responses in the resulting simulation. Their experiments reveal that geometry with small structural details can be excluded from acoustic modelling, maintaining the perceived quality, and this acts as a motivating basis for this future iterative study.

C. Geometry Reduction

Determining space subdivision through the resolution factor P^3 of the *geometry reduction* stage has an effect on the volume reconstruction and generation of image patches, which directly map to acoustic materials. Larger resolutions require more marching cubes, causing the number of orthographic projections from surfaces to increase, resulting in a higher number of forward passes through the feature extractor, finally resulting in increased computational overhead. In order to maintain perceptual accuracy and produce plausible acoustic simulations whilst minimising the spatial resolution to optimise execution times, further work would require subjective evaluations to derive optimal spatial resolution across varying scene geometry.

D. Objective Evaluation

The most noticeable differences between renders with predicted and manually tagged materials are due to different decays of acoustic energy. By considering spectrograms of generated RIRs (3), the different acoustic materials influence the decay of energy over time. As a result, there are errors relative to reverberation, definition and clarity; see Table I.

VI. Conclusions

We presented a novel pipeline for acoustic rendering that is able to capture acoustic material characteristics of space around a listener using computer vision paradigms. These predicted acoustic material characteristics are used to generate input for sound propagation methods, producing plausible acoustic simulations. We developed a proof-of-concept prototype of an automated pipeline, executed as offline procedures

mostly implemented as CPU programs, demonstrating the generation of RIRs that can be used in downstream convolution auralisations in real-time audio engines to propagate audio from virtual sound sources in simulated environments. The automated mapping between visual appearances and acoustic characteristics directly applies to extended reality platforms where the virtual environment is incrementally reconstructed as a listener explores their surroundings, enabling sound rendering to produce plausible acoustic simulations removing human experts from the scene authoring process.

REFERENCES

[1] E. Lakka, A. G. Malamos, K. G. Pavlakis, and J. A. Ware, "Spatial sound rendering–a survey," *IJIMAI*, vol. 5, no. 3, pp. 33–45, 2018.

[2] H. Hacihabiboglu, E. De Sena, Z. Cvetkovic, J. Johnston, and J. O. Smith III, "Perceptual spatial audio recording, simulation, and rendering: An overview of spatial-audio techniques based on psychoacoustics," *IEEE Signal Processing Magazine*, vol. 34, no. 3, pp. 36–54, 2017.

[3] J. L. Rubio-Tamayo, M. Gertrudix Barrio, and F. García García, "Immersive environments and virtual reality: Systematic review and advances in communication, interaction and simulation," *Multimodal Technologies and Interaction*, vol. 1, no. 4, p. 21, 2017.

[4] P. Larsson, D. Vastfjall, and M. Kleiner, "Better presence and performance in virtual environments by improved binaural sound rendering," *Journal of the Audio Engineering Society*, june 2002.

[5] C. Chen, U. Jain, C. Schissler, S. V. A. Gari, Z. Al-Halah, V. K. Ithapu, P. Robinson, and K. Grauman, "Soundspaces: Audio-visual navigation in 3d environments," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*. Springer, 2020, pp. 17–36.

[6] M. T. Taylor, A. Chandak, L. Antani, and D. Manocha, "Resound: interactive sound rendering for dynamic virtual environments," in *Proceedings of the 17th ACM international conference on Multimedia*, 2009, pp. 271–280.

[7] D. Manocha and M. C. Lin, "Interactive sound rendering," in *2009 11th IEEE International Conference on Computer-Aided Design and Computer Graphics*. IEEE, 2009, pp. 19–26.

[8] Z. Tang, H.-Y. Meng, and D. Manocha, "Learning acoustic scattering fields for dynamic interactive sound propagation," *arXiv preprint arXiv:2010.04865*, 2020.

[9] N. Singh, J. Mentch, J. Ng, M. Beveridge, and I. Drori, "Image2reverb: Cross-modal reverb impulse response synthesis," *arXiv preprint arXiv:2103.14201*, 2021.

[10] Z. Tang, N. J. Bryan, D. Li, T. R. Langlois, and D. Manocha, "Scene-aware audio rendering via deep acoustic analysis," *IEEE transactions on visualization and computer graphics*, vol. 26, no. 5, pp. 1991–2001, 2020.

[11] H. Kim, L. Remaggi, P. J. Jackson, and A. Hilton, "Immersive spatial audio reproduction for vr/ar using room acoustic modelling from 360 images," in *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. IEEE, 2019, pp. 120–126.

[12] A. Reilly and D. McGrath, "Convolution processing for realistic reverberation," in *Audio Engineering Society Convention 98*. Audio Engineering Society, 1995.

[13] N. Raghuvanshi and J. Snyder, "Parametric wave field coding for precomputed sound propagation," *ACM Transactions on Graphics (TOG)*, vol. 33, no. 4, pp. 1–11, 2014.

[14] B. Hamilton and S. Bilbao, "Fdtd methods for 3-d room acoustics simulation with high-order accuracy in space and time," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 11, pp. 2112–2124, 2017.

[15] R. Mehra, N. Raghuvanshi, L. Savioja, M. C. Lin, and D. Manocha, "An efficient gpu-based time domain solver for the acoustic wave equation," *Applied Acoustics*, vol. 73, no. 2, pp. 83–94, 2012. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0003682X11001605

[16] L. Savioja and U. P. Svensson, "Overview of geometrical room acoustic modeling techniques," *The Journal of the Acoustical Society of America*, vol. 138, no. 2, pp. 708–730, 2015.

[17] H. Kuttruff, *Room acoustics*. Crc Press, 2016.

[18] A. A. de Lima, T. de M. Prego, S. L. Netto, B. Lee, A. Said, R. W. Schafer, T. Kalker, and M. Fozunbal, "Feature analysis for quality assessment of reverberated speech," in *2009 IEEE International Workshop on Multimedia Signal Processing*, 2009, pp. 1–5.

[19] C. Schissler and D. Manocha, "Interactive sound propagation and rendering for large multi-source scenes," *ACM Trans. Graph.*, vol. 36, no. 4, Sep. 2016. [Online]. Available: https://doi.org/10.1145/3072959.2943779

[20] M. Vorländer, "Simulation of the transient and steady-state sound propagation in rooms using a new combined ray-tracing/image-source algorithm," *The Journal of the Acoustical Society of America*, vol. 86, no. 1, pp. 172–178, 1989. [Online]. Available: https://doi.org/10.1121/1.398336

[21] S. Siltanen, T. Lokki, L. Savioja, and C. Lynge Christensen, "Geometry reduction in room acoustics modeling," *Acta Acustica united with Acustica*, vol. 94, no. 3, pp. 410–418, 2008.

[22] M. Colombo, A. Dolhasz, and C. Harvey, "A computer vision inspired automatic acoustic material tagging system for virtual environments," in *2020 IEEE Conference on Games (CoG)*. IEEE, 2020, pp. 736–739.

[23] C. Schissler, C. Loftin, and D. Manocha, "Acoustic classification and optimization for multi-modal rendering of real-world scenes," *IEEE transactions on visualization and computer graphics*, vol. 24, no. 3, pp. 1246–1259, 2017.

[24] S. Bell, P. Upchurch, N. Snavely, and K. Bala, "Opensurfaces: A richly annotated catalog of surface appearance," *ACM Transactions on graphics (TOG)*, vol. 32, no. 4, pp. 1–17, 2013.

[25] A. Chang, A. Dai, T. Funkhouser, M. Halber, M. Niessner, M. Savva, S. Song, A. Zeng, and Y. Zhang, "Matterport3d: Learning from rgb-d data in indoor environments," *arXiv preprint arXiv:1709.06158*, 2017.

[26] M. Colombo, A. Dolhasz, and C. Harvey, "A texture superpixel approach to semantic material classification for acoustic geometry tagging," in *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021, pp. 1–7.

[27] U. Gaur and B. S. Manjunath, "Superpixel embedding network," *IEEE Transactions on Image Processing*, vol. 29, pp. 3199–3212, 2020.

[28] L. Savioja, J. Huopaniemi, T. Lokki, and R. Väänänen, "Creating interactive virtual acoustic environments," *Journal of the Audio Engineering Society*, vol. 47, no. 9, pp. 675–705, 1999.

[29] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.

[30] P. Bourke, "Polygonising a scalar field," 1994.

[31] E. Lengyel, *Foundations of Game Engine Development: Volume 2: Rendering*. Terathon Software, 2019.

[32] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[33] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[34] E. A. Habets, "Room impulse response generator," *Technische Universiteit Eindhoven, Tech. Rep*, vol. 2, no. 2.4, p. 1, 2006.

[35] S. W. Smith *et al.*, *The scientist and engineer's guide to digital signal processing*. California Technical Pub. San Diego, 1997.

[36] J. M. Kates and E. J. Brandewie, "Adding air absorption to simulated room acoustic models," *The Journal of the Acoustical Society of America*, vol. 148, no. 5, pp. EL408–EL413, 2020.

[37] S. Adavanne, A. Politis, and T. Virtanen, "TUT Sound Events 2018 - Circular array, Anechoic and Synthetic Impulse Response Dataset," Apr. 2018. [Online]. Available: https://doi.org/10.5281/zenodo.1237752

[38] P. Manocha, Z. Jin, R. Zhang, and A. Finkelstein, "Cdpam: Contrastive learning for perceptual audio similarity," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 196–200.

[39] D. Diaz-Guerra, A. Miguel, and J. R. Beltran, "gpurir: A python library for room impulse response simulation with gpu acceleration," *Multimedia Tools and Applications*, vol. 80, no. 4, pp. 5653–5671, 2021.

[40] M. Taylor, A. Chandak, Q. Mo, C. Lauterbach, C. Schissler, and D. Manocha, "Guided multiview ray tracing for fast auralization," *IEEE Transactions on Visualization and Computer Graphics*, vol. 18, no. 11, pp. 1797–1810, 2012.

[41] S. Pelzer and M. Vorländer, "Frequency-and time-dependent geometry for real-time auralizations," in *Proceedings of 20th International Congress on Acoustics, ICA*, 2010, pp. 1–7.