

# Immersive Audio-Visual Scene Reproduction using Semantic Scene Reconstruction from 360 Cameras

Hansung Kim · Luca Remaggi · Aloisio Dourado · Teofilo de Campos · Philip JB Jackson · Adrian Hilton

Received: date / Accepted: date

**Abstract** As personalised immersive display systems have been intensely explored in Virtual Reality (VR), plausible 3D audio corresponding to the visual content is required to provide more realistic experiences to users. It is well known that spatial audio synchronised with visual information improves a sense of immersion but limited research progress has been achieved in immersive audio-visual content production and reproduction. In this paper, we propose an end-to-end pipeline to simultaneously reconstruct 3D geometry and acoustic properties of the environment from a pair of omni-directional panoramic images. A semantic scene reconstruction and completion method using a deep convolutional neural network is proposed to estimate the complete semantic scene geometry in order to adopt spatial audio reproduction to the scene. Experiments provide objective and subjective evaluations of the proposed pipeline for plausible audio-visual VR reproduction of real scenes.

**Keywords** Audio-Visual scene reproduction · Scene understanding · 3D reconstruction and completion · Spatial audio

---

H. Kim  
ECS, University of Southampton, UK  
E-mail: h.kim@soton.ac.uk

L. Remaggi  
Creative Labs UK  
E-mail: luca.remaggi@cle.creative.com

A. Dourado and T. de Campos  
University of Brasilia, Brazil  
E-mail: t.decampos@oxfordalumni.org

P.J.B. Jackson and A. Hilton  
CVSSP, University of Surrey, UK  
E-mail: a.hilton@surrey.ac.uk

## 1 Introduction

In Virtual Reality (VR) systems, personalised audio-visual experiences are one of the most important issues to improve the sense of presence because human perception relies on audio and visual cues to understand and interact with the environment [34, 49]. However, most existing approaches have primarily focused on a single modality. Recent research combines audio and vision into systems to enable semantic scene understanding and human interaction [73, 84].

A full 3D reproduction of a real space in a virtual environment allows users to experience the space remotely. It can be widely applied to various fields such as teleconferencing [54], education [9, 63], health care [50], entertainment [57] and media production [18, 40]. However, research has mainly focused on improving the visual side of scene reconstruction. In immersive VR systems, users do not perceive the scene as realistic if sound is not matched with the visual cues [27]. For example, sounds should be provided with the correct early-reflections and reverberation effect which the user expects from the visual scene [4]. This also allows correct perception of distance to the sound source [58]. Many studies performed evaluations of reconstructed 3D visual scenes [44, 56], but the quality of the audio has not been considered. Some researches investigated virtual reality auralizations [64, 71] but they were not directly synchronised with real visual scenes. Recent research has investigated scene-aware spatial audio reproduction in 2D panoramic video rendering using a mono-channel microphone/speaker pair recording [51] and self-supervised deep learning [60].

This paper provides a practical solution to capture room structure and acoustic properties allowing spatial audio to be adapted to the 3D model of a room environment and listener location to give a plausible rendering to improve immersion. We propose a full 3D reconstruction pipeline with acoustic

property estimation from a pair of off-the-shelf consumer omni-directional (360°) camera captures of indoor scenes. Two 360° panoramic images are used to reconstruct a complete semantic scene geometry model and render the spatial audio in the environment. A preliminary version of the approach presented in this paper previously appeared at a conference [43], which estimates an acoustic room model from 360° images. However, the previous work approximates room geometry with large cuboids without any detail and the pipeline is inefficient as it is composed of two separate processes: 2D object recognition and 3D geometry reconstruction. The object labels inferred from the 2D image are projected to the reconstructed 3D model to segment semantic objects. Our proposed pipeline is an integrated 3D pipeline that is more accurate and works significantly faster than [43] in building detailed 3D geometry with semantic information. It also reproduces more plausible spatial audio in the reconstructed scene models. The main contributions and advantages of this paper over the preliminary work are:

- Complete audio-visual VR scene reconstruction system using a pair of consumer 360° image captures.
- Semantic scene reconstruction and completion from 360° stereo images taking advantage of existing standard RGB-D datasets for network training.
- Real-time user interactive audio-visual VR scene rendering with spatial audio.
- Comprehensive objective and subjective evaluations of estimated room geometry and acoustics.

## 2 Background and motivation

### 2.1 3D modelling from images

Indoor 3D geometry modelling from images has been extensively researched. A huge number of Multi-View Stereo (MVS) [23], Simultaneous Localization and Mapping (SLAM) [15] and Structure from Motion (SfM) [10] algorithms using multiple photos/videos have been developed. Low-cost RGB-D (RGB + depth) cameras have also made a great impact on real-time indoor scene reconstruction [59]. However, due to the limited field-of-view (FoV) of imaging sensors, these methods require multiple images or video streams to cover the whole scene.

360° cameras (also known as panoramic or omnidirectional cameras) which capture all directions at the same time using fish-eye or wide FoV lenses have been recently introduced to our daily life. These off-the-shelf low-cost 360° cameras used in many practical applications [6, 61] can provide a good solution for this coverage problem. Song et al. proposed a SfM method from a 360° camera [77]. Im et al. proposed a dense depth map estimation pipeline using a narrow-baseline video clip captured by a 360° camera [38].

We also proposed scene reconstruction methods using stereo 360° images from various types of 360° cameras [41, 43]. We followed this stereo-based method to acquire a depth map for images captured by 360° cameras as this allows simple set up and capture processes, as well as dynamic scene captures.

### 2.2 3D semantic scene reconstruction and completion

Despite remarkable progress in image-based 3D reconstruction, the incomplete reconstruction problem caused by occlusions due to the physical limitations of the capture process still remains. 3D semantic scene reconstruction and completion was initially proposed by Song et al. [78]. From a given single RGB-D image, they build a semantically labelled 3D voxel structure including occluded and non-surface regions based on a fully convolutional neural network (CNN) with 3D dilated convolutions jointly trained for semantic object segmentation and scene completion. This work also introduced the use of Flipped Truncated Signed Distance Function (F-TSDF) to encode the depth map projected to 3D before feeding it to the 3D CNN. Zhang et al. [88] proposed to use Spatial Group Convolutions to reduce the amount of computational resources for network training. Liu et al. [53] improved it using a two step training protocol composed of a 2D semantic segmentation CNN and a 3D semantic scene completion CNN. Kim et al. [45] proposed a 3-D scene graph for a semantic representation of rooms. We proposed EdgeNet [20], an integrated architecture using edge information detected from the corresponding RGB image. EdgeNet was designed for normal perspective images. In this paper we extend it to 360° images for whole scene reconstruction and completion. One problem of 360° scene reconstruction and completion is the lack of ground-truth 360° RGB-D data for training. There are a few 3D 360° datasets such as Stanford 2D-3D-Semantics dataset [2] and Matterport 3D [16], but the number of scenes provided by those datasets are not enough for training CNN architectures. On the other hand, there are abundant normal RGB-D datasets available with annotated ground-truth for training. In this research, we propose to decompose the 360° view into 8 overlapping views to benefit from existing RGB-D datasets for training and enable complete 360° scene reconstruction.

### 2.3 Acoustic modelling for spatial audio rendering

Various methods have been developed to describe the characteristics of room acoustics through sets of parameters, which enable reproduction of real-world spatial audio effects in virtual scenes [62, 65, 86]. The parameters are typically extracted from measured acoustical Room Impulse Responses (RIRs) [62, 83]. For VR scenes, RIRs can be synthesised

from the room geometry [42]. However, modelling room acoustics with an RIR is still incomplete as the RIR is only valid for a single source-receiver configuration and it is impractical to measure or update all RIRs according to the changes of geometry or source/user positions in interactive rendering environments. It also takes time and resources to set up a bulky loudspeaker, microphone and audio system to make measurements. Recently, a few vision-based approaches to estimate room acoustics for spatial audio rendering have been proposed [48]. 3D models with material information allow the emulation of real world acoustics [37]. Li et al. proposed scene-aware spatial audio reproduction from a single video recording [51] but it was only for 2D 360° video rendering. Schissler et al. built a dense 3D geometry using a SfM method from RGB-D image frames and estimate acoustic material properties for sound rendering using a CNN [74], but it requires an RGB-D video stream of the static scene to cover a complete structure estimation.

Many audio tool kits have been recently developed to render spatial audio. G’Audio provides an object-based spatial audio plug-in for a 3D environment but supports limited platforms [25]. Wwise Spatial Audio plug-in supports a wide range of VR platforms including Unreal and Unity to efficiently model sound propagation in a given 3D space [46]. Google Resonance [28] and Steam Audio [17] also provide free open-source plug-ins for immersive spatial audio rendering which can handle multiple occlusions, reflection, reverb and HRTF effects in a VR environment. We use Google Resonance and Steam Audio to embed estimated acoustic parameters and render spatial audio in the reconstructed 3D semantic scene models. However, our final experiments are not intended as comparison between the Google Resonance and Steam VR performance. Instead, we employed these two tools since they are two of the most relevant ones, and they provide free open-source plug-ins for immersive spatial audio rendering, which enabled us to create our end-to-end pipeline. Detailed features of these two toolkits are introduced in Section 3.4.

### 3 Proposed Pipeline

#### 3.1 System Overview

The ultimate goal of this research is to develop a practical system for reproduction of visually and acoustically plausible VR scenes from a simple capture of indoor scenes. Figure 1 shows the flow of the proposed pipeline.

A full surrounding scene is captured as a vertical stereo image pair with 360° cameras. This pair of images is used for depth estimation of the scene using stereo matching and depth map enhancement. From the estimated depth map, an initial voxel-based structure is generated and partitioned into eight overlapped parts. The partitioned voxel structures are

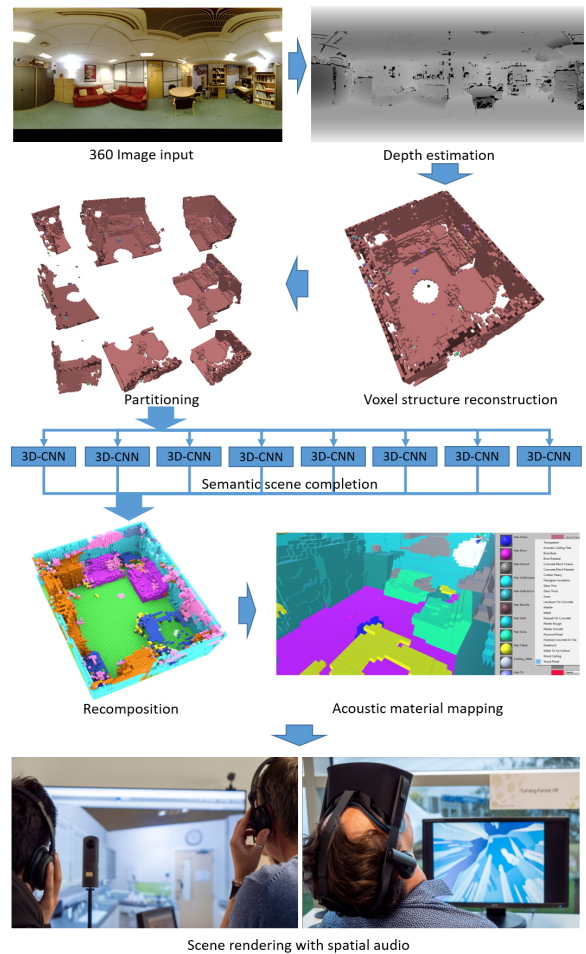


Fig. 1 Overview of the proposed pipeline

individually submitted to the EdgeNet with the corresponding edge maps for semantic segmentation and scene completion. The outputs of EdgeNet are recomposed into the whole scene structure and acoustic property parameters for the classified objects are assigned from the acoustic material list. Finally, the reconstructed audio-visual VR scene is rendered by setting sound source and player models on a Unity [85] VR platform.

#### 3.2 Capture system and depth map generation

One of requirements for practical audio-visual applications is a simple and quick capture/recording process of the real scene. Nowadays, inexpensive off-the-shelf 360° cameras which produce high quality of scenes capture are widely available [29, 39]. Ricoh Theta cameras [67] were used in our system as they provide accurately rectified equi-rectangular photos from two fisheye lens and also support first-order Ambisonics (B-format) audio recording.

Two Ricoh Theta cameras were set on vertically aligned mounts to capture the scene with full panoramic texture and

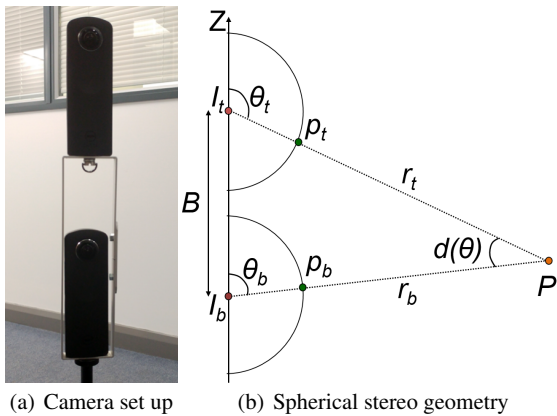


Fig. 2 Spherical stereo geometry with 360° cameras

to extract depth information as shown in Fig. 2 (a). From the pair of vertical stereo 360° images, depth of the scene is estimated by dense correspondence matching [41]. According to the spherical stereo geometry in Fig. 2 (b), depth information can be recovered from pixel disparity and stereo camera baseline distance  $B$  without any camera calibration as column and row positions in an equi-rectangular image are directly mapped to the azimuth and elevation angles respectively in the 3D spherical coordinate system. If the angular disparity of two matching points ( $\theta_t(p)$ ,  $\theta_b(p')$ ) between two images is given as  $d(\theta(p)) = \theta_t(p) - \theta_b(p')$ , the distance (depth)  $r_t(p)$  from the top camera to the real 3D scene point is calculated by triangulation as:

$$r_t(p) = B / \left( \frac{\sin \theta_t(p)}{\tan(\theta_t(p) - d(\theta(p)))} - \cos \theta_t(p) \right). \quad (1)$$

Any correspondence matching algorithm can be used for the proposed pipeline. We used a simple feature-based bi-directional block matching method in our experiments. However, scene depth recovery using correspondence matching from stereoscopic images is subject to noisy depth from matching failure and incomplete scene depth due to occlusions.

In order to increase the prediction performance in semantic scene completion, the estimated depth map is enhanced under the assumption that most objects in an indoor scene are piecewise-planar [32] and edge information is a distinguishing feature for reliable stereo matching, providing good depth estimates on their neighbourhood [12]. The Canny edge detector [26] is applied to the image to detect candidate regions for piecewise-planar regions. Using the dilated edges as a mask, the most reliable depth estimations are extracted from the original depth map. Vertical edges are eliminated from the mask as they are parallel to the epipolar lines and do not contribute to the stereo matching performance in the given vertical stereo camera set up. Coherent regions with similar colours are searched by a simple flood fill approach to detect featureless planar surfaces like single colored walls and table tops. Planes are fitted to those

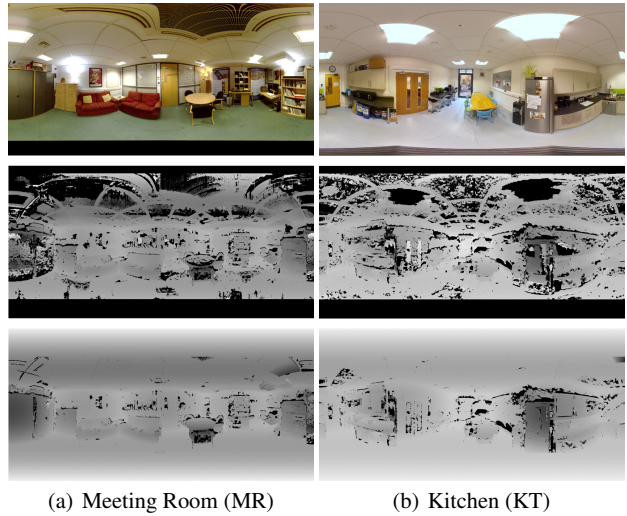


Fig. 3 Depth Enhancement (Top: Original Top image; Middle: Estimated depth map; Bottom: Enhanced depth map)

regions using RANSAC [22] to eliminate noise from false stereo matching. The original depth information is replaced by the depth of the plane if the plane is closely aligned to any principal axis. The original depth information is kept for non-orthogonal plane regions.

Figure 3 shows examples of the depth enhancement results. The cabinets in the left part of the MR scene have serious depth errors due to the vertical stripes on the surface, but most errors are eliminated by the enhancement step. The depth errors on the shiny surface of the fridge are also corrected in the KT scene. The ceiling and floor regions in both scenes have wide erroneous areas due to the featureless surface or saturated lighting, but they were approximated to smooth planes.

### 3.3 Semantic scene reconstruction and completion

A 3D voxel structure of all visible surfaces can be reconstructed by projecting points in the estimated depth map to a 3D space, but this structure is incomplete due to invisible regions in the scene. In this section we propose a semantic scene reconstruction and completion for 360° scenes simultaneously filling occluded areas and segmenting the structure into semantic parts. This work is based on the semantic scene completion using EdgeNet for a normal perspective (narrow FoV) RGB-D image [20]. We extended this EdgeNet to 360° scene completion and understanding while taking advantage of existing standard RGB-D datasets for network training.

The voxel structure is partitioned into 8 partially overlapped views from the center of the scene. The FoV of the partitioned view is set to 45° to match to the FoV of the standard RGB-D sensor used for network training, and the

viewpoint of the partition in each direction is positioned 1.7m behind the original camera position to get overlapped coverage to compensate the boundary regions in each partition. Each partition is individually submitted to the semantic scene completion network for prediction. In our experiments, the whole 3D volume is represented by  $480 \times 480 \times 144$  voxels with 0.02m voxel size, and each partition size is set as  $240 \times 240 \times 144$  voxels. The resolution can be increased according to the memory allowance.

### 3.3.1 Training datasets

The 3D CNN architecture used in our pipeline is trained on the SUNCG training set [78] and fine-tuned on the NYU depth v2 (NYU-v2) training set [75]. The SUNCG dataset has 140K RGB-D views extracted from 45K synthetic scenes with corresponding depth maps and ground truth. The NYU-v2 dataset includes 464 real scene depth and RGB images (795 views for training and 654 views for testing) captured by a Kinect sensor. We generated ground truth by voxelizing the 3D mesh annotations from [31] and mapped object categories based on [33] to label occupied voxels with semantic object classes.

### 3.3.2 EdgeNet for semantic scene completion

Figure 4 illustrates the 3D CNN architecture to build a complete and air-tight 3D structure with object labels from depth and edge maps. This has been inspired by the U-Net design [69], and differs from other Semantic Scene Completion approaches by fusing both depth and edges after encoding using F-TSDF [79]. The edge volume is generated from the edge map as a voxel structure with the same dimension as the depth volume. The activation function of EdgeNet is a Softmax and each voxel of the output volume contains the predicted probabilities of the 12 classes used for training. The output resolution for each partition is  $60 \times 36 \times 60$  voxels.

The One Cycle Learning policy [76] combined with Curriculum Learning [8] and Simulated Annealing [1] is used for the training stage. In the fine-tuning stage, the network is initialised with the SUNCG parameters and tuned using standard training with a Stochastic Gradient Descent (SGD) optimizer with a learning rate of 0.01 and decay weight of 0.0005. The training time was about 4 days on SUNCG and 6 hours on NYU on an Nvidia GTX 1080 Ti GPU.

### 3.3.3 Recomposition

The output of the EdgeNet-based semantic scene completion architecture is 8 object-labeled 3D volumes, that have overlaps at their boundaries with their neighbours. In order to

combine the output partitions into one complete scene structure, a simple strategy of “summing *a posteriori* probability” proposed by Kittler et al. [47] is applied for each class over all classifier outputs.

All output partitions are located at their original positions and all voxels in the ranges are checked if they belong to certain partitions or not. If a given voxel is not covered by a certain partition, *a posteriori* probabilities for all classes for that voxel and partition are set as 0 (out of FoV). Otherwise, the sum of the *a posteriori* probabilities for all classes for that voxel and classifier is set as 1. For a voxel with *a posteriori* probability  $P_{ij}$  for class  $i$  predicted by a classifier  $j$ , the sum of the probabilities for class  $i$  over all classifiers  $n$  is given by:

$$S_i = \sum_{j=1}^n P_{ij}. \quad (2)$$

The winning class  $C$  for this voxel is:

$$C = \arg \max_i (S_i). \quad (3)$$

## 3.4 Room acoustics and VR scene reproduction

This semantic scene structure is directly imported to Unity to simulate room acoustics. We initially considered two well-known tools to simulate spatial audio in the Unity engine: Google Resonance [28] and Steam Audio [17]. They both implement their spatial auralisation by employing binaural RIRs (BRIRs) over virtual loudspeakers.

Google Resonance provides 22 types of acoustic materials, and Steam Audio provides 11 preset acoustic materials and 1 custom material property setting. Both Google Resonance and Steam Audio calculate the early reflections using Head Related Transfer Functions (HRTFs), belonging to the closest Direction of Arrival (DOA) estimated via ray tracing. The employed HRTFs are obtained through interpolation: the available HRTFs corresponding to the DOAs which are the nearest to the reflection DOA are used to perform HRTF interpolation. We used the initial HRTF datasets built in the Steam Audio Unity Plug-in<sup>1</sup>.

The main difference about the way Google Resonance and Steam VR render spatial sound is that Google Resonance uses a two-step approach: first it places the sources onto a high-order Ambisonics field and then it reproduces the obtained field through virtual loudspeakers [68], [30]. Instead, Steam follows a single-step approach: it generates the BRIRs related to virtual loudspeakers directly. Steam Audio was developed to generate an accurate acoustic simulation while Google Resonance aims to bring the spatial

<sup>1</sup> [https://valvesoftware.github.io/steam-audio/doc/phonon\\_unity.html](https://valvesoftware.github.io/steam-audio/doc/phonon_unity.html)



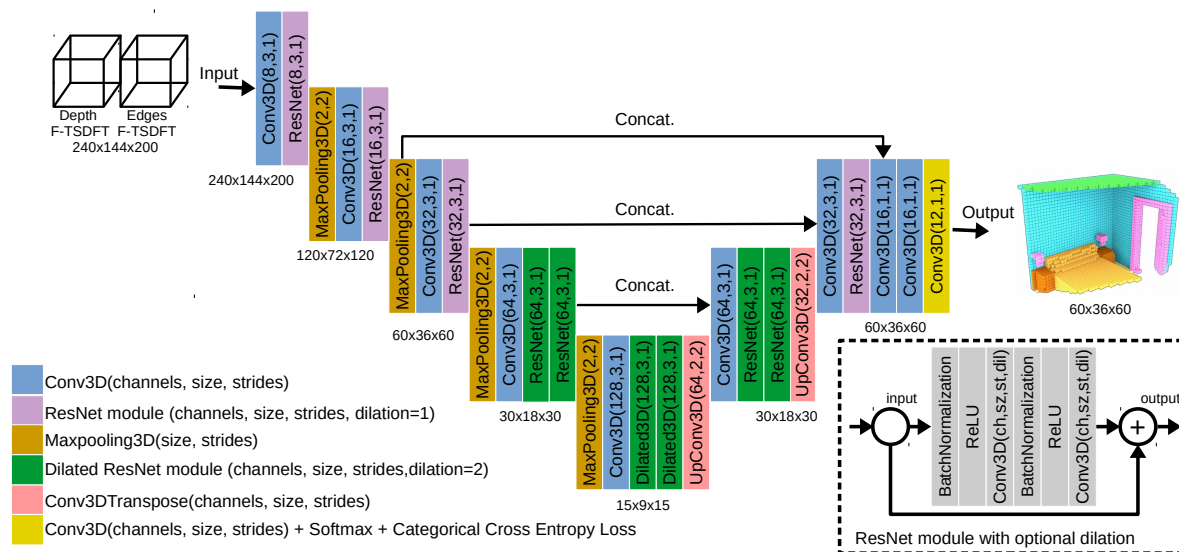


Fig. 4 U-shaped architecture for complete 3D semantic scene reconstruction

audio experience to mobile devices reducing the computational complexity.

We found that Google Resonance has several **limitations in our system implementation**: (a) Google Resonance does not work with voxel-based structure but only with mesh-based surface structure; (b) **the audio quality suffers when rendering sound with simple frequency content, such as a sine wave beep or a swept sine signal, which is generally used for RIR measurement but not recommended for spatialization**<sup>2</sup>. Therefore, we used Steam Audio for our final implementation though we still included results with Google Resonance in the experiment to verify the performances of both tool kits.

In order to render spatial audio from the estimated acoustic properties of the reconstructed 3D models, we map the object labels to the acoustically closest material types in the provided audio package as Table 1 as an approach to estimate acoustic properties of materials from a visual input.

Although measuring RIRs in real environments is well-established [80], extracting RIR information from VR environments has not previously been explored. Therefore, we treat the virtual environment as a real one to measure BRIRs, and emulate virtual binaural microphones and omnidirectional sound sources in the reconstructed virtual environments, to record sounds. The general swept-sine method [21] is employed to calculate RIRs for Steam Audio, and an anechoic gun-shot (normalised in the time domain) [19] is used for Google Resonance. The BRIRs for reconstructed 3D environments are obtained by recording the responses at the same positions as the Ground-truth BRIRs measured in the real environment.

Table 1 Object-Material matching table.

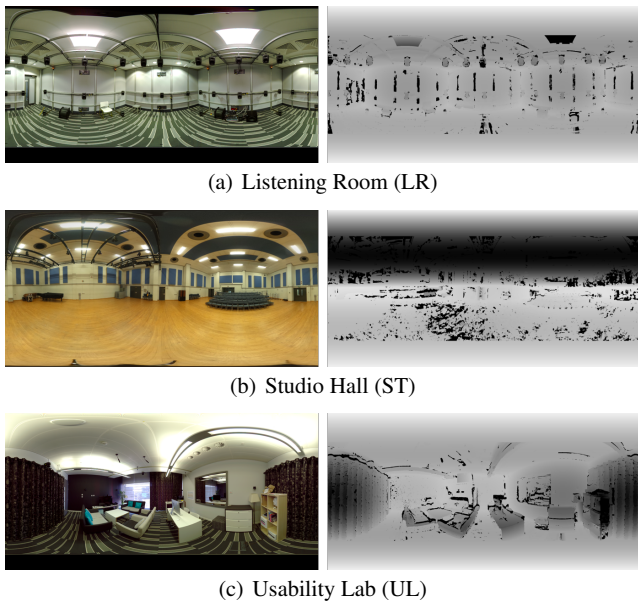
Object	Material - Google	Material - Steam
Empty	Transparent	Transparent
Ceiling	Wood ceiling	Wood
Floor	Curtain heavy	Carpet
Wall	Plaster smooth	Plaster
Window	Thick Glass	Glass
Bed	Heavy curtain	Carpet
Sofa	Heavy curtain	Carpet
Chair	Plywood panel	Wood
Table	Plywood panel	Wood
TV	Thick Glass	Glass
Furniture	Plywood panel	Wood
Object	Metal	Metal

Finally, a virtual camera and audio sources are placed in the VR scene to render the reconstructed scene with spatial audio. The reproduced VR scene is rendered with real-time interaction on a VR headset or desktop applications. In our experiments, HTC VIVE Pro [36], a VR headset playing binaural spatial audio over headphones is used.

## 4 Experiments

In this section, we present our experiments to evaluate the quality of visual geometry and acoustics reproduced by the proposed pipeline. In the evaluation of immersive spatial audio, both “authenticity” and “plausibility” of rendered sound should be considered. Authenticity measures how identical the generated sound is to the ground-truth sound [11], while plausibility is subjectively judged by the listener with his/her inner reference (listener’s expectation) [52]. Authenticity of the rendered sound can be evaluated by comparing acoustic parameters of the rendered sound with those of the ground-truth sounds (objective evaluation). Plausibility can be eval-

<sup>2</sup> <https://resonance-audio.github.io/resonance-audio/develop/design-tips.html>



**Fig. 5** Dataset used in the experiments (Left: Captured Top image; Right: Enhanced depth map)

uated by user studies. Both objective and subjective evaluations have been carried out in this study.

The proposed pipeline has been tested on five different rooms with various sizes and materials: Meeting Room (MR), Usability Lab (UL), Kitchen (KT), Listening Room (LR) and Studio Hall (ST). The MR and UL scenes are typical living room environments. KT is a long and narrow room with kitchen equipment. LR is an acoustically controlled experimental room and ST is a large hall. Each scene was captured as a 360° vertical stereo pair. The MR and KT sets are shown in Fig. 3, and the other datasets with their enhanced depth maps are in Fig. 5.

All data sets, audio sources, results and supplementary video in this section are available at:

<http://3dkim.com/research/VR/index.html>

#### 4.1 Semantic scene reconstruction

Full 3D structures of the scenes with semantic object labels were reconstructed through the proposed depth estimation and semantic scene completion process. The results were compared with the block-based scene reconstruction method (Kim19) [43] which employs two separate process of 2D semantic segmentation by SegNet [3] and room modelling by cuboid fitting. Figure 6 shows the reconstructed scene models with semantic object labels indexed by the colour code in Fig. 8. The first row in Fig. 6 shows the initial voxel clouds generated by the estimated depth maps in the 3D space with a voxel size of 0.02m, before encoding the volumes with F-TSDF and submitting them to the proposed networks. It is observed that the initial structures are incomplete due to the

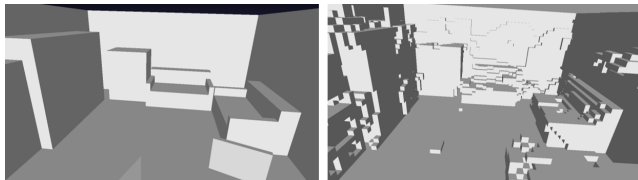
occlusions and erroneous depth estimation. They also have two large holes at the epipoles of vertical stereo (under and over the camera location). The second and third rows visualise the outputs by [43] and the proposed pipeline, respectively. In the outputs of [43], the walls were indexed as Floor as the whole room layout was represented as one cuboid, but we assigned correct materials for walls and ceilings in acoustic material mapping. In the results of [43], the side cabinets in the MR scene do not adjoin the wall as the side parts of those cabinets are not visible in the captured image. It also missed the large table in the KT scene and produced many redundant objects from scattered loud speakers in the LR scene. Overall, the proposed 360 semantic scene reconstruction method produced more objects correctly located in the scene with geometrical details such as the tea tables between two sofas in the MR scene, the main table in the KT scene and the curtain on the wall in the UL scene.

It is difficult to quantitatively evaluate the reconstruction performance for individual objects in the rooms because ground-truth models are not available. For a preliminary evaluation, we made a CG model by manual measurements for the MR scene. For the KT scene, a LiDAR scan data was available, but the LiDAR appears also to fail with transparent surfaces like the doors and windows. There is a hole on the floor too. Though the reference models are still incomplete due to approximation and occlusion, it is observed that the estimated model by the proposed method generates approximate geometry of the main objects in the scene in Fig. 6. We also evaluated the room dimensions against the manually measured room layouts in Table 2. Both [43] and the proposed method built relatively accurate room layouts. The proposed method shows slightly better estimations for room width and length, but [43] was a bit better in the room height estimation because the depth for the whole ceilings and floors were inferred from limited number of features in the proposed method. The estimation errors were relatively large for the ST scene because the accuracy of depth estimation for spherical stereo is inverse-proportional to the distance. We used a fixed baseline distance for all scenes but this can be improved by setting the baseline distance larger for a large scene. The height of the ST scene was estimated incorrectly (14.46% of error) due to the uneven ceiling with rails and panels in the scene. [43] showed large errors in width and length for the UL scene due to the large window and mirror in the scene but the proposed method accurately matched the room layouts.

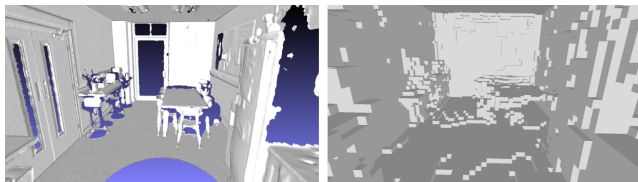
The semantic scene reconstruction was run on a GeForce GTX TITAN X GPU with 12GB memory and the whole process took around 2 mins per dataset. This is much faster than [43] which requires two separate process for 8 mins (3 mins of 2D semantic segmentation and 5 mins of 3D geometry reconstruction).

**Table 2** Comparison of reconstruction errors in room dimensions. The bold figures represent the minimum error in each dimension.

Data	Ground-truth	Kim19 [43]		Proposed	
	Dimension (m)	Dimension (m)	Err in Dim (%)	Dimension (m)	Err in Dim (%)
MR	5.61×4.28×2.33	5.52×4.35×2.36	(1.60, 1.64, <b>1.29</b> )	5.54×4.24×2.40	( <b>1.25</b> , <b>0.93</b> , 3.00)
KT	6.64×3.46×2.67	6.95×3.41×2.70	(4.67, <b>1.45</b> , 1.12)	6.42×3.52×2.68	( <b>3.31</b> , 1.73, <b>0.37</b> )
LR	5.64×5.05×2.90	5.77×5.17×2.98	( <b>2.30</b> , 2.38, <b>2.76</b> )	5.88×5.02×2.78	(4.26, <b>0.59</b> , 4.14)
ST	17.08×14.55×6.50	16.53×14.87×5.70	(3.22, <b>2.20</b> , <b>12.31</b> )	17.54×15.46×5.56	( <b>2.69</b> , 6.25, 14.46)
UL	5.57×5.20×2.91	5.92×4.95×2.95	(6.28, 4.81, <b>1.37</b> )	5.52×5.22×3.00	( <b>0.90</b> , <b>0.38</b> , 3.09)



(a) MR (Left: Approximate CG model, Right: Estimated geometry)



(b) KT (Left: LiDAR scanning, Right: Estimated geometry)

**Fig. 6** Comparison of geometry reconstruction

## 4.2 Room acoustics evaluation

For objective evaluation of the sound rendered in the reconstructed 3D models, RIRs simulated in the reconstructed models are compared with the ground-truth RIRs measured in the real environments. The ground-truth RIRs were recorded by using the swept-sine method [21] with a sampling frequency of 48 kHz, employing a Genelec 8020B speakers as sound source and a Soundfield MK5 B-format microphone. For the objective evaluation, the W-channel (i.e. omni) of the Soundfield microphone was used. In fact, only the timbrally-encoded spatial effects of the room response were evaluated through the objective metrics, leaving the complete spatial evaluation to the subjective tests, which were run on a desktop application.

### 4.2.1 Evaluation Metrics

To evaluate the quality of the acoustics reproduced in the reconstructed environments, we analysed the Early Decay Time (EDT) and Reverberation Time (RT60) of the generated sounds, as objective measures of their early reflections and late reverberation, respectively. EDT is a good metric to evaluate the acoustics from adjacent reflectors that is subjectively important, by considering the energy carried by the early reflections [13, 70]. On the other hand, RT60 relates to the average absorption, location of room boundaries and size of the room, describing the reverberation from a phys-

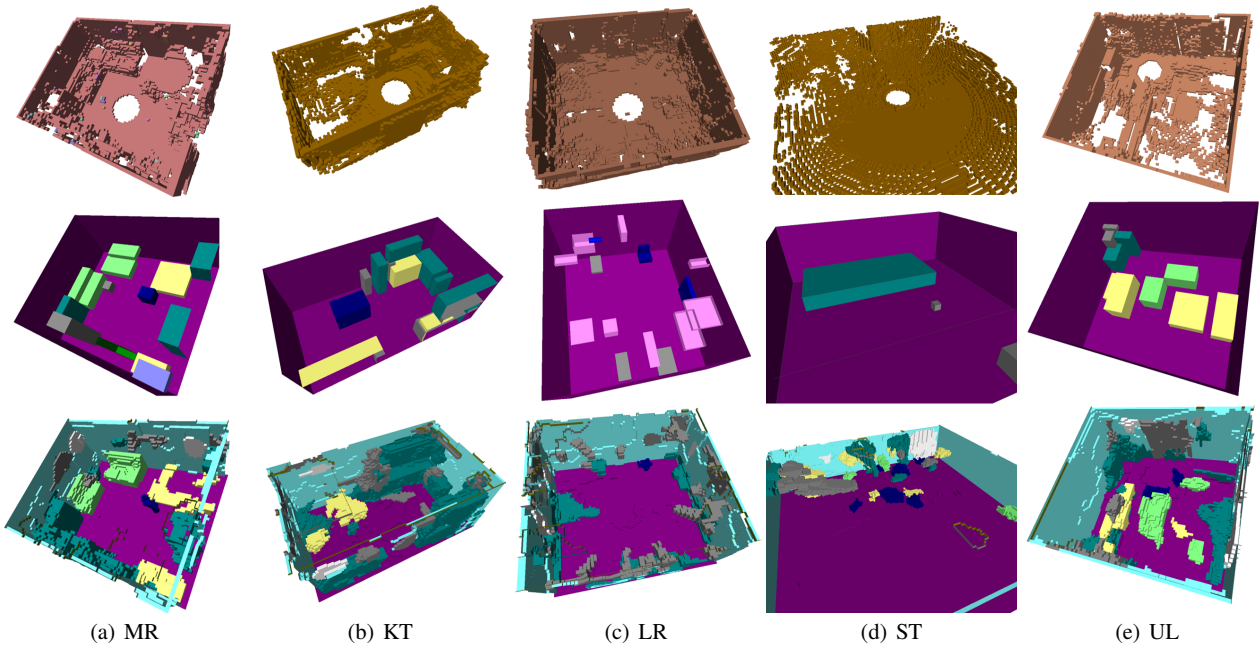
ical point of view [13, 70]. EDT is calculated as six times the time required for the energy to decay 10 dB after the direct sound [7]. RT60 is measured as the time for the energy to decay 60 dB. The average values over the 6 octave bands between 250 Hz and 8 kHz are reported for both EDT and RT60 in this research.

To understand the perceptual meaning of the observed errors in the EDT and RT60 values, we defined their Just Noticeable Differences (JNDs). The thresholds of JND were chosen as 20 % for RT60 [55] and 5 % for EDT [87] as suggested in the literature. However, it is important to note that the same literature describes variable percentages depending on the type of sounds. For instance, in [55], JNDs for RT60 were found up to about 30 % for musical signals. It is also important to remark that, for media or entertainment applications, authenticity is not the benchmark to target. It is widely recognised that sound plausibility is more important [11, 52]. To the best of our knowledge, no threshold that defines the plausibility limens for the object metrics employed has been identified in the literature. Previous studies typically focused on determining plausibility by observing the overall sound perception, without distinguishing between the perception of early reflections and late reverberation [58]. Furthermore, in the presence of visual stimuli, the perceptual differences between real and synthetic acoustic environment are not as strictly defined as they are for unimodal scenarios [5]. In this paper, we employed JNDs to be coherent with the available literature though JNDs typically refer to authenticity in audio-only scenarios, the strictest case.

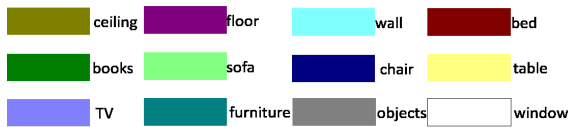
### 4.2.2 Evaluation results

The EDTs and RT60s of rendered RIRs were compared against the ground-truth data measured in the real environments and visualised in Fig. 9 and 10, respectively. The proposed method was compared along with three other models. First, “Kim19” is a state-of-the-art method [43] illustrated in Fig. 6. Second, “Empty” is an empty shoe box model which represents only room boundaries with correct materials to verify the role of objects within the scene for sound rendering. Third, for “Wrong”, the geometry was reconstructed based on the proposed method but incorrect acoustic material classification was manually assigned to the three largest objects in the scene to prove the importance of accurate material estimation in sound rendering. There is no carpet in the KT scene,

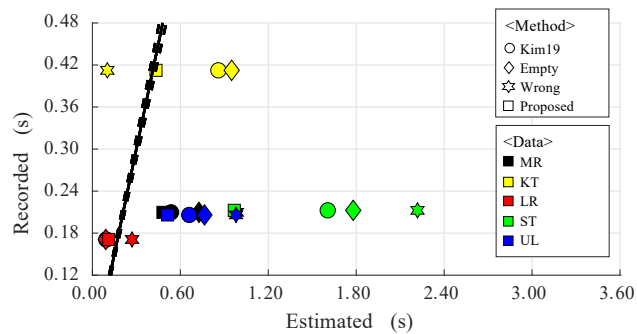




**Fig. 7** Semantic room reconstruction results (Top: Initial voxel cloud, Middle: Reconstruction by [43], Bottom: Proposed method, The color index is defined as in Fig. 8



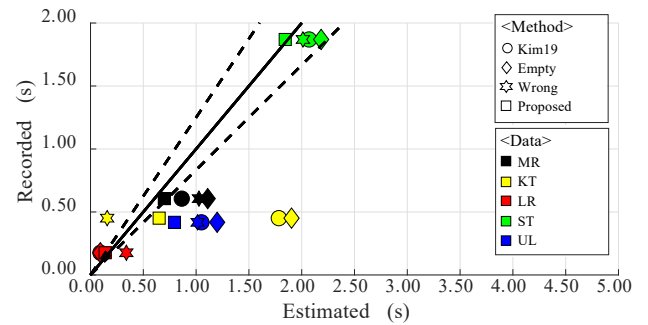
**Fig. 8** Colour index for semantic objects linked to the materials in Table 1



**Fig. 9** EDTs for the five rooms, related to the estimated RIRs in VR environment. The dashed lines show the JND limit of 5% [87].

and other scenes have less hard surfaces. Therefore, we assigned Carpet as the wrong material to the KT, and Glass to the other scene to maximise the difference of acoustic effect.

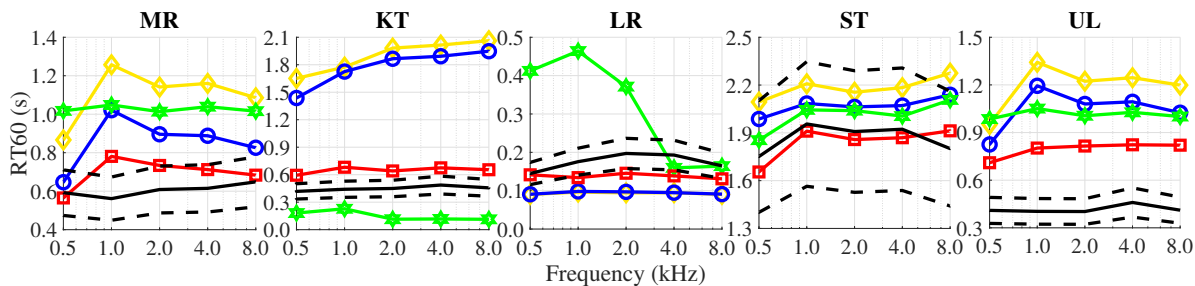
In Fig. 9, the EDT results show that the proposed method outperforms all other methods. In particular, for the KT and LR scenes, the estimated values fall close to the JND band, which means very small perceptual differences from the recorded wrong materials comparing with the “Kim19 [43]” models



**Fig. 10** RT60s for the five rooms, related to the estimated RIRs in VR environment. The dashed lines show the JND limit of 20% [55].

which show the strong relationship between interior objects (both geometry and material) and room acoustics.

Regarding the RT60 results in Fig. 10, there are similar trends to the ones observed in the EDT evaluations. The proposed method shows the best performance among the tested methods. For the ST and MR scenes, RT60s by the proposed method are inside the JND band, which means that the proposed method has recreated authentic representations of the room reverberation. One interesting observation in this RT60 test is that “Wrong” performs slightly better than “Kim19 [43]” for KT, ST and UL scenes (i.e. 3 rooms out of 5). This suggests the geometrical details reconstructed by the proposed method can improve RT60 in spite of the which have very simple box-shaped geometry.



**Fig. 11** RT60s over the different frequency bands for the five rooms. (Black line: ground-truth, Dashed black lines: JND, Blue circle: Kim19 [43]. Yellow diamond: Empty, Green star: Wrong, Red square: Proposed)

Figure 11 shows the RT60s depending on frequency. The proposed method generates signals closest to the recorded ground-truth, for every tested frequency. As also previously observed in Fig. 10, the general trend is to have “Empty” to be the worst and “Wrong” is then the second best for KT, ST and UL scenes. The largest error for the proposed method seems to appear in UL because the curtain which actually absorbs sound in the UL scene has been classified as ‘objects’ in the proposed method.

### 4.3 Subjective evaluation

The aim of this experiment is to evaluate the plausibility in terms of the perceived spatial impression and quality of the sound rendered by the proposed pipeline. The evaluation test was performed by twenty participants with normal hearing between the ages of 22 and 45, made up of 12 males and 8 females. None of them was experienced in 3D audio. Wired Bose QC25 headphones were used for the sound reproduction and a 24” monitor screen to observe the related room images in a 360° rendering mode for the test. All tests were embedded into the MUSHRA interfaces<sup>3</sup>, developed in Max MSP.

#### 4.3.1 Listening Test Setup

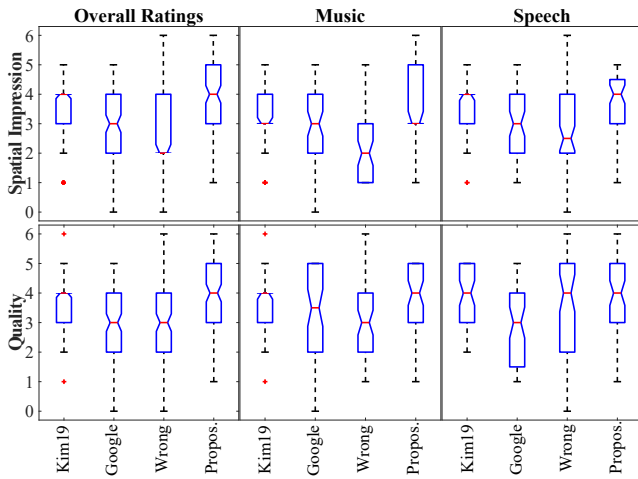
Four rooms in the experimental dataset were selected for subjective evaluation: MR, KT, ST and UL. The Listening Room (LR) was excluded because people cannot expect sound rendered in the room with acoustically insulated walls from the given image. Two original sound sources were rendered in the VR environments: an anechoic speech source from the TIMIT dataset [24] and a clarinet sound in an anechoic chamber from the OpenAirLib library [14].

In order to evaluate subjective attributes of the reproduced sound quality, two factors were tested as proposed in [35]: “spatial impression” and “overall quality”. The participants were presented with a PC interface having a 360°

image viewer and an audio player with corresponding audio tracks to rank each stimulus against the attribute, within a range of integer numbers from 0 to 6. We did not provide the full 3D VR scene to avoid the visual cues influencing perception of the acoustic cues. The experiment was made up of eight sessions, the combination of four rooms and two audio sources, and two questions were given to the participants to rate the stimulus: (1) “How much the spatial impression matches what you expect from the given room image”; (2) “How natural the generated sound is against any noise or distortion”.

Six samples were provided per session on the MUSHRA interface: two reference samples (‘Low’ and ‘High’) as anchors and randomly assigned four test samples (‘A’-‘D’) generated by different methods and audio tools. The participants were free to listen to reference and test samples by clicking buttons on the interface to rate the test samples. The low and high references were generated as anchors to help the listeners. The ‘Low’ reference samples are the original sound sources recorded in an anechoic environment which does not include any of reverberation. The ‘High’ references were binaural sounds generated from B-format (i.e., first-order Ambisonics) recordings. This conversion was done using the NoiseMakers Ambi-Head plug-in, in Reaper. The first test sample “Kim19” was generated by Steam Audio from the model estimated by the state-of-the-art algorithm [43] used in the previous Sections 4.1 and 4.2. The second sample “Google” was generated by the Google Resonance package to evaluate its audio rendering performances. The third sample “Wrong” is the one with wrong acoustic materials introduced in the previous Section 4.2 and rendered by Steam Audio. Finally, the “Proposed” sample was rendered by Steam Audio with the semantic 3D model reconstructed by the proposed pipeline. We restricted the total number of samples to six per session and excluded other variable combinations to prevent the listeners getting confused or tired. The four test samples were randomly shuffled to the buttons ‘A’ to ‘D’ in every session so that the participant could not find any consistency from the order.

<sup>3</sup> <https://github.com/IoSR-Surrey/MUSHRA-MaxMSP>



**Fig. 12** Subjective scores for spatial impression (top three) and overall quality (bottom three). On each box, the horizontal red line represents the median of the distribution, the bottom and top edges of the boxes are the 25<sup>th</sup> and 75<sup>th</sup> percentiles, respectively. The whiskers show the most extreme non-outlier samples, whereas the outliers are the red stars.

### 4.3.2 Spatial Impression

In this test, two reference samples were provided: the ‘High’ one, the binaural sound obtained from B-format, was rated as 5; whereas, the ‘Low’ one, the anechoic recordings, rated as 1. The participants were asked to rate four samples within the score range 0-6. (They could give even 0 or 6 if they felt any sample was worse or better than the references).

The spatial impression results are visualised in the top row of Fig. 12. The overall rating shows the proposed method to provide a better spatial impression compared to the other three. The proposed method shows the same median values (i.e. 4 out of 6) as “Kim19” [43], but looking at their 75<sup>th</sup> percentile, the results related to the proposed method are more stretched towards values greater than 4, as opposed to the results related to “Kim19” [43]. It is also important to show that participants clearly perceived the lowest spatial impression when the wrong materials were given (“Wrong”), again demonstrating the importance of correct object and material recognition.

The second and third figures in the top row show the split between the two content types: music and speech. The trend observed in the overall score is mainly given by the speech results. For the music, all methods seem to provide similar spatial impression, but the data distribution is also stretched towards higher values for the proposed method, as suggested by the 75<sup>th</sup> percentile.

These plots could, mistakenly, lead to the conclusion that the four tested methods produce results having similar statistics. Therefore, we have run further statistical analysis over the results. In Table 3, we report the results of the one-way ANOVA test [81], which aims to identify whether

the different methods produced a statistically significant effect on the results. The test results confirm this hypothesis. This can be observed by looking at the p-values that are always (much) below 5%: with a confidence greater than 95%, the distributions are statistically different. In Table 4, we also reported the results’ means, and p-values obtained by running t-tests between each pair of methods [82]. The reported mean values confirm what was already observed: for the spatial impression, the proposed method performs the best, on average, followed by (in order): “Kim19”, “Google” and “Wrong”. Regarding the paired t-tests, we first compared “Kim19” to “Google”, “Wrong” and “Proposed”; then “Google” to “Wrong” and “Proposed”; finally, “Wrong” to “Proposed”. We can see that the differences related to “Proposed” are statistically significant when directly compared to the distributions of the results of “Google” and “Wrong”. This confirms that our pipeline composed by room geometry estimation and Steam VR is able to render an acoustic field that gives a better spatial impression than the other analysed pipelines. Nevertheless, these results also show that it is not possible to claim, with 95% confidence, that the “Proposed” results are significantly better than our previous work in “Kim19”.

### 4.3.3 Overall Quality

In this test, the participants evaluated the stimulus only in terms of sound quality, scoring them between 0 and 6. The same eight sessions were tested with the same four test samples without references.

The results are reported in the bottom row of Fig. 12. The observed trend is similar to the one discussed for the spatial impression. In general, the participants seem to prefer the quality provided by the proposed method. Nonetheless, Kim19 [43] also gives good performance, with the median being the same as the new method but its 75<sup>th</sup> percentile is distributed towards lower values. The sounds reproduced by Google Resonance seem to be perceived as having low quality. This difference in quality is due to the fundamental differences between Google Resonance and Steam Audio as discussed in Section Google Resonance employs an Ambisonic field as intermediate step to create virtual loudspeakers, whereas Steam directly generates the BRIRs. For an infinite-order Ambisonic rendering via virtual loudspeakers, the direct sound and early reflections would be sharp in terms of their spatial image and the full room effect would be conveyed via the surrounding virtual loudspeakers. However, the limited-order Ambisonic decoding (they suggest to use 3rd-order to achieve the highest fidelity) not only blurs the spatial image of the direct sound and early reflections but causes temporal smearing of the impulse response at the ears as a result of panning combined with variation in the propagation time to each ear (relative to the calibrated/aligned

**Table 3** F-Values and related P-values of the one-way ANOVA tests among the different methods’ ratings, for overall rating, single music and single speech results.  $H = 1$  means that the test rejects the null hypothesis of the groups belonging to normal distributions with equal means (95 % of confidence), also denoted by \*.

	Overall Ratings			Music			Speech		
	H	F-Value	P-Value	H	F-Value	P-Value	H	F-Value	P-Value
Spatial Impression	1	17.71	< 0.1 %*	1	10.20	< 0.1 %*	1	8.35	< 0.1 %*
Overall Quality	1	7.73	< 0.1 %*	0	2.17	9.2 %	1	8.60	< 0.1 %*

**Table 4** Means of the results and the scores obtained by paired t-tests between the different methods’ ratings, for overall rating, single music and single speech results.  $H = 1$  means that the pair of results are statistically different (95 % of confidence). The first four rows’ label SI stands for Spatial Impression, whereas the last four rows’ label OQ stands for Overall Quality.

	Overall Ratings				Music				Speech			
	Kim19	Google	Wrong	Prop.	Kim19	Google	Wrong	Prop.	Kim19	Google	Wrong	Prop.
SI - Means	3.41	2.97	2.57	<b>3.63</b>	3.25	2.95	2.30	<b>3.54</b>	3.57	3.00	2.84	<b>3.73</b>
SI - t-test Kim19	-	$H = 1$	$H = 1$	$H = 0$	-	$H = 0$	$H = 1$	$H = 0$	-	$H = 1$	$H = 1$	$H = 0$
SI - t-test Google	-	-	$H = 1$	$H = 1$	-	-	$H = 1$	$H = 1$	-	-	$H = 0$	$H = 1$
SI - t-test Wrong	-	-	-	$H = 1$	-	-	-	$H = 1$	-	-	-	$H = 1$
OQ - Means	3.71	3.05	3.21	<b>3.72</b>	3.54	3.34	3.07	<b>3.64</b>	<b>3.88</b>	2.77	3.56	3.79
OQ - t-test Kim19	-	$H = 1$	$H = 1$	$H = 0$	-	$H = 0$	$H = 1$	$H = 0$	-	$H = 1$	$H = 1$	$H = 0$
OQ - t-test Google	-	-	$H = 0$	$H = 1$	-	-	$H = 0$	$H = 0$	-	-	$H = 1$	$H = 1$
OQ - t-test Wrong	-	-	-	$H = 1$	-	-	-	$H = 1$	-	-	-	$H = 0$

delay to the center of the listener). Using HRIRs for the final stage of convolution to render the virtual loudspeakers to binaural cannot simultaneously compensate for all active directions of arrival. Since Steam has a single stage method, it avoids this problem.

Therefore, Steam seems to be a better option when high sound quality is required, while Google Resonance would be more appropriate for light and fast rendering applications [66]. Moreover, Google Resonance usually performs better when near-field sounds are reproduced, and when source directionality is important. In fact, Steam only allows omnidirectional sources and its HRTF interpolation algorithm seems less accurate in near-field scenarios. Nevertheless, these important features are not related to the aim of our study here. Therefore, their effect on the spatial sound quality did not emerge from our listening tests.

The ANOVA test results without any multiple hypothesis adjustment, e.g., Bonferroni [72, 81], for the overall quality are in Table 3. For the overall ratings and speech the p-values are always below 1 %, demonstrating that, with a confidence greater than 95 %, the four method result distributions are statistically different. Nevertheless, this cannot be claimed for music results. In that case, a p-value of 9.2 % cannot reject (with a 95 % of confidence) the null hypothesis of the results’ normal distributions having same means. Looking at the means in Table 4, we can see that overall, and in particular with music, the proposed approach performs the best. However, with speech, our previous “Kim19” is the best, with “Proposed” ranked second. Regarding the paired t-tests, similar to the spatial impression, the results related to “Proposed” are typically statistically significant with respect

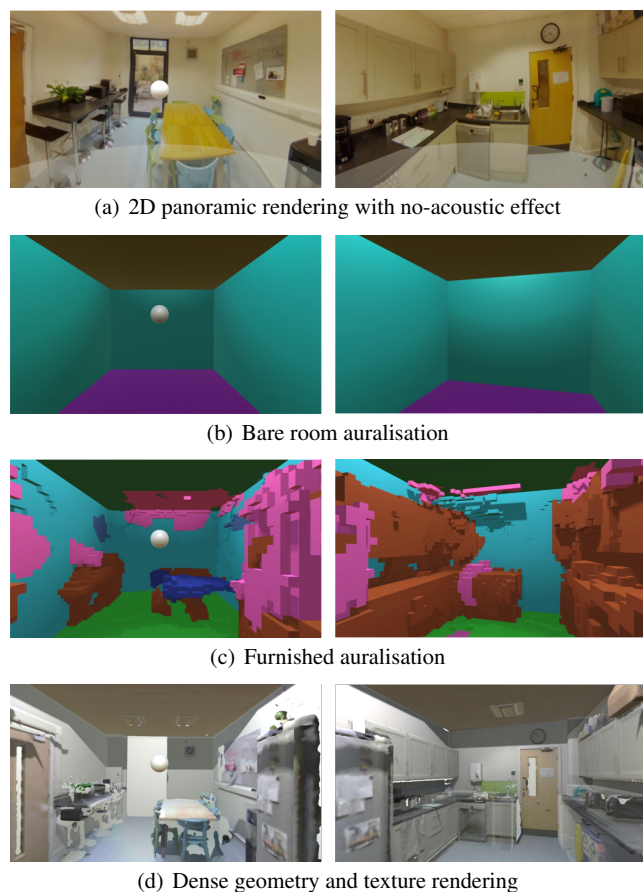
to “Google” and “Wrong”. Nevertheless, it is not possible to claim, with 95% confidence, that “Proposed” and “Kim19” give a statistically significant improvement.

However, by looking at these results together with the spatial impression’s, we can conclude that, overall, “Proposed” is the best pipeline among the tested ones, with “Kim19” being very close to it, in particular providing the best overall quality with speech.

#### 4.4 Interactive VR scene rendering

The final VR scenes with spatial audio reproduced by the proposed pipeline were played on a VR headset, HTC VIVE Pro [36] with real-time user interaction. In this implementation, users can freely navigate with 6 degrees of freedom (DoF) in the virtual scene by their own movements or using the Vive controller. This implementation provides four different simulation modes shown in Fig. 13: (a) 360° panoramic scene with the original sound without any acoustic effect; (b) Bare-room auralisation only with 3D room layout without any object; (c) Furnished auralisation with full 3D room structure and acoustic properties (Proposed method); (d) Dense geometry and texture visualisation with the same auralisation as (c) (Proposed method). The white sphere in the scenes shows the location of the sound source. In mode (d), dense geometry and texture have been overlaid on the transparent voxel structure to increase the level of visual immersion. The spatial audio was still rendered with the reconstructed voxel structure, as in mode (c). We did not use the dense geometry for spatial audio rendering in order to keep low computational complexity for real-time rendering. In Fig. 13 (a), the





**Fig. 13** Interactive real-time VR scene rendering with spatial audio

360° textures are not exactly matched to the rendered 3D geometry because the panoramic image was simply projected on a sphere as a 2D texture.

It is hard to quantitatively evaluate plausibility of audio-visual VR content because the perception of the acoustic environment is influenced by visual cues [5]. We demonstrated this interactive audio-visual system with the VR headset in several public events and received informal verbal feedback. The sounds rendered by the proposed pipeline were demonstrated and compared with the original source and the sound rendered in the empty room. Users consistently reported a higher sense of immersion when the spatial audio was given together with dense 3D geometry and texture.

## 5 Conclusion

We proposed a practical solution to reproduce plausible audio-visual VR scenes from 360° images allowing spatial audio to be adapted to the virtual model of a room environment. The first part of the proposed pipeline is a vision-based method to estimate the complete room model with semantic information. A voxel-based 3D model of the scene is reconstructed and completed with semantic labels using

an ensemble of 3D CNNs trained using normal perspective image datasets. This information is used to generate spatial audio on Unity with audio tool kits, allowing perceptually plausible sound for the scene.

The reproduced room geometry and spatial audio were evaluated against actual data measured and recorded in the original rooms. The proposed method obtained much faster semantic scene reconstruction with geometric details and achieved better agreement between the real and simulated acoustics than the state-of-the-art algorithm through objective and subjective evaluations.

Future work will look at robust material recognition in the acoustic room modelling to replace the current object-to-material mapping method as the current object categories are not enough to represent the full range of acoustic properties. For example, a solid “Wall” recognised from visual cues cannot tell if it is a concrete wall or light partition wall. A “Sofa” cannot tell if the material is fabric or leather. A multi-modal sensory approach, e.g., combining audio-visual sensors, can be a solution to create a stronger relationship between objects and their acoustic properties.

**Acknowledgements** This work was supported by the EPSRC Programme Grant S3A: Future Spatial Audio for an Immersive Listener Experience at Home (EP/L000539/1), the BBC as part of the BBC Audio Research Partnership, and Audio-Visual Media Research Platform (EP/P022529/1). Details about the data underlying this work are available from: <http://dx.doi.org/10.15126/surreydata.00812228>.

## References

1. Aarts, E., Korst, J.: *Simulated Annealing and Boltzmann Machines: A Stochastic Approach to Combinatorial Optimization and Neural Computing*. John Wiley & Sons, Inc., New York, NY, USA (1989)
2. Armeni, I., Sener, O., Zamir, A.R., Jiang, H., Brilakis, I., Fischer, M., Savarese, S.: 3D semantic parsing of large-scale indoor spaces. In: *Proc. CVPR*, pp. 1534–1543 (2016)
3. Badrinarayanan, V., Kendall, A., Cipolla, R.: Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2017)
4. Bailey, W., Fazenda, B.M.: The effect of reverberation and audio spatialization on egocentric distance estimation of objects in stereoscopic virtual reality. *J. Acoustical Society of America* **141**(5), 3510 (2017)
5. Bailey, W., Fazenda, B.M.: The effect of visual cues and binaural rendering method on plausibility in virtual environments. In: *Proc. of the 144th AES Convention* (Milan, Italy, 2018)
6. Barazzetti, L., Previtali, M., Roncoroni, F.: Can we use low-cost 360 degree cameras to create accurate 3d models? *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* **XLII-2**, 69–75 (2018)
7. Barron, M.: Interpretation of early decay times in concert auditoria. *Acta Acustica United with Acustica* **81**(4), 320–331(12) (1995)
8. Bengio, Y., Louradour, J., Collobert, R., Weston, J.: Curriculum learning. In: *Proc. ICML*, pp. 41–48 (2009)



9. Bhamu, P.R.K.S., Hariharasubramanian, V., Mythili, O.P., Ramachandran, M.: Users' domain knowledge prediction in e-learning with speech-interfaced augmented and virtual reality contents. *Virtual Reality* **24**, 163–173 (2017)
10. Bianco, S., Ciocca, G., Marelli, D.: Evaluating the performance of structure from motion pipelines. *Journal of Imaging* **4**(8) (2018)
11. Blauert, J.: *Communication Acoustics*. Springer-Verlag Berlin Heidelberg (2005)
12. Bleyer, M., Breiteneder, C.: Stereo matching - state-of-the-art and research challenges. In: *Advanced Topics in Computer Vision*, pp. 143–179 (2013)
13. Bradley, J.S.: Review of objective room acoustics measures and future needs. *Applied Acoustics* **72**(10), 713–720 (2011)
14. Brown, K., Paradis, M., Murphy, D.: Openairlib: A javascript library for the acoustics of spaces. In: *Audio Engineering Society Convention* 142 (2017)
15. Cadena, C., Carlone, L., Carrillo, H., Latif, Y., Scaramuzza, D., Neira, J., Reid, L., Leonard, J.: Past, present, and future of simultaneous localization and mapping: Towards the robust-perception age. *IEEE Transactions on Robotics* **32**(6), 1309–1332 (2016)
16. Chang, A., Dai, A., Funkhouser, T., Halber, M., Niessner, M., Savva, M., Song, S., Zeng, A., Zhang, Y.: Matterport3D: Learning from RGB-D data in indoor environments. In: *Proc. 3DV* (2017)
17. Corporation, V.: Steam audio. <https://valvesoftware.github.io/steam-audio/> (2021)
18. Cosker, D., Eisert, P., Grau, O., Hancock, P.J.B., McKinnell, J., Ong, E.: Applications of face analysis and modeling in media production. *IEEE MultiMedia* **20**(4), 18–27 (2013)
19. Cox, T.: Gun shot in anechoic chamber. Freesound: <https://freesound.org/people/acs272/sounds/210766/> (2013)
20. Dourado, A., de Campos, T.E., Kim, H., Hilton, A.: EdgeNet: Semantic scene completion from rgb-d images. In: *Proc. ICPR* (2021)
21. Farina, A.: Simultaneous measurement of impulse response and distortion with a swept-sine technique. In: *Proc. of the AES Convention* (2000)
22. Fischler, M.A., Bolles, R.C.: Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM* **24**(6), 381–395 (1981)
23. Furukawa, Y., Hernández, C.: Multi-view stereo: A tutorial. *Found. Trends. Comput. Graph. Vis.* **9**(1-2), 1–148 (2015)
24. Garofolo, J.S., Lamel, L.F., Fisher, W.M., Fiscus, J.G., Pallet, D.S., Dahlgren, N.L.: DARPA TIMIT acoustic phonetic continuous speech corpus CDROM. Tech. rep., NIST Interagency (1993)
25. Gaudio: Gaudio vr audio. <https://gaudiolab.com/solution-ar-vr-and-immersive/> (2021)
26. Gonzalez, R., Woods, R.: *Digital Image Processing*. Pearson (2017)
27. Gonzalez-Franco, M., Lanier, J.: Model of illusions and virtual reality. *Frontiers in Psychology* **8**(1), 1125 (2017)
28. Google: Google resonance audio. <https://resonance-audio.github.io/resonance-audio/> (2021)
29. GoPro: Gopro fusion. <https://shop.gopro.com/EMEA/cameras/fusion/CHDHZ-103-master.html> (2019)
30. Gorzel, M., Allen, A., Kelly, I., Gungormusler, A., Kammerl, J., Yeh, H., Boland, F.: Efficient encoding and decoding of binaural sound with resonance audio. In: *Proc. of the AES Conference on Immersive and Interactive Audio* (York, UK, 2019)
31. Guo, R., Zou, C., Hoiem, D.: Predicting complete 3D models of indoor scenes. *CoRR abs/1504.02437* (2015). URL <http://arxiv.org/abs/1504.02437>
32. Gupta, A., Efros, A.A., Hebert, M.: Blocks world revisited: Image understanding using qualitative geometry and mechanics. In: *Proc. ECCV* (2010)
33. Handa, A., Patraucean, V., Badrinarayanan, V., Stent, S., Cipolla, R.: SceneNet: Understanding real world indoor scenes with synthetic data. *CoRR abs/1511.07041* (2015). URL <http://arxiv.org/abs/1511.07041>
34. Hicks, M., Nichols, S., O'Malley, C.: Comparing the roles of 3d representations in audio and audio-visual collaborations. *Virtual Reality* **7**, 148–163 (2004)
35. Hoeg, W., Christensen, L., Walker, R.: Subjective assessment of audio quality - the means and methods within the EBU. Tech. rep., EBU Technical Review (1997)
36. HTC: Vive pro. <https://www.vive.com/uk/product/vive-pro-full-kit/> (2018)
37. Hulusic, V., Harvey, C., Debattista, K., Tsingos, N., Walker, S., Howard, D., Chalmers, A.: Acoustic rendering and auditory-visual cross-modal perception and interaction. *Journal of Computer Graphics Forum* **31**(1), 102–131 (2012)
38. Im, S., Ha, H., Rameau, F., Jeon, H.G., Choe, G., Kweon, I.: All-around depth from small motion with a spherical panoramic camera. In: *Proc. ECCV*, vol. 9907 (2016)
39. Insta360: Insta360 one x. <https://www.insta360.com/product/insta360-onex> (2019)
40. Kim, H., Guillemaut, J.Y., Takai, T., Sarim, M., Hilton, A.: Outdoor dynamic 3d scene reconstruction. *IEEE Transactions on Circuits and Systems for Video Technology* **22**(11), 1611–1622 (2012)
41. Kim, H., Hilton, A.: 3D scene reconstruction from multiple spherical stereo pairs. *International Journal of Computer Vision* **104**(1), 94–116 (2013)
42. Kim, H., Hughes, R.J., Remaggi, L., Jackson, P.J.B., Hilton, A., Cox, T.J., Shirley, B.: Acoustic room modelling using a spherical camera for reverberant spatial audio objects. In: *Proc. of the 142th AES Convention* (2017)
43. Kim, H., Remaggi, L., Jackson, P.J., Hilton, A.: Immersive spatial audio reproduction for vr/ar using room acoustic modelling from 360 images. In: *Proc. IEEE VR* (2019)
44. Kim, H.G., Lim, H., Ro, Y.M.: Deep virtual reality image quality assessment with human perception guider for omnidirectional image. *IEEE Transactions on Circuits and Systems for Video Technology* **30**(4), 917–928 (2020)
45. Kim, U., Park, J., Song, T., Kim, J.: 3-d scene graph: A sparse and semantic representation of physical environments for intelligent agents. *IEEE Transactions on Cybernetics* **50**(12), 4921–4933 (2020)
46. Kinetic, A.: Wwise spatial audio. <https://www.audiokinetic.com/products/wwise-spatial-audio/> (2021)
47. Kittler, J., Hatef, M., Duin, R.P.W., Matas, J.: On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **20**(3), 226–239 (1998)
48. Kon, H., Koike, H.: Deep neural networks for cross-modal estimations of acoustic reverberation characteristics from two-dimensional images. In: *Proc. of the 144th AES Convention* (Milan, Italy, 2018)
49. Larsson, P., Våljamäe, A., Västfjäll, D., Tajadura-Jiménez, A., Kleiner, M.: Auditory-Induced Presence in Mixed Reality Environments and Related Technology (2010)
50. Laver, K.E., George, S., Thomas, S., Deutsch, J.E., Crotty, M.: Virtual reality for stroke rehabilitation. *The Cochrane Collaboration* **2015**(2), 1–27 (2015)
51. Li, D., Langlois, T.R., Zheng, C.: Scene-aware audio for 360° videos. *ACM Transactions on Graphics* **37**(4) (2018)
52. Lindau, A., Weinzierl, S.: Assessing the plausibility of virtual acoustic environments. *Acta Acustica united with Acustica* **98**(5), 804–810 (2012)
53. Liu, S., HU, Y., Zeng, Y., Tang, Q., Jin, B., Han, Y., Li, X.: See and think: Disentangling semantic scene completion. In: S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, R. Garnett (eds.) *Proc. NIPS*, pp. 263–274 (2018)

54. Mekuria, R., Blom, K., Cesar, P.: Design, implementation, and evaluation of a point cloud codec for tele-immersive video. *IEEE Transactions on Circuits and Systems for Video Technology* **27**(4), 828–842 (2017)
55. Meng, Z., Zhao, F., He, M.: The just noticeable difference of noise length and reverberation perception. In: *Proc. of the International Symposium on Communications and Information Technologies* (Bangkok, Thailand, 2006)
56. Menzies, R., Rogers, S.J., Phillips, A.M., Chiarovano, E., Waele, C., Verstraten, F., MacDougall, H.: An objective measure for the visual fidelity of virtual reality and the risks of falls in a virtual environment. *Virtual Reality* **20**, 173–181 (2016)
57. Narayanan, S., Polys, N., Bukvic, I.: Cinemacraft: exploring fidelity cues in collaborative virtual world interactions. *Virtual Reality* **24**, 53–73 (2020)
58. Neidhardt, A., Tommy, A.I., Pereppadan, A.D.: Plausibility of an interactive approaching motion towards a virtual sound source based on simplified BRIR sets. In: *Proc. of the 144th AES Convention* (Milan, Italy, 2018)
59. Newcombe, R., Izadi, S., Hilliges, O., Molyneaux, D., Kim, D., Davison, A., Kohli, P., Shotton, J., Hodges, S., Fitzgibbon, A.: Kinectfusion: Real-time dense surface mapping and tracking. In: *Proceedings of ISMAR* (2011)
60. Pedro Morgado Nuno Vasconcelos, T.L., Wang, O.: Self-supervised generation of spatial audio for 360° video. *In: Proc. NIPS* (2018)
61. Peng, X., Bennamoun, M., Wang, Q., Ma, Q., Xu, Z.: A low-cost implementation of a 360° vision distributed aperture system. *IEEE Transactions on Circuits and Systems for Video Technology* **25**(2), 225–238 (2015)
62. Politis, A., Tervo, S., Lokki, T., Pulkki, V.: Parametric multidirectional decomposition of microphone recordings for broadband high-order ambisonic encoding. In: *Proc. of the 144th AES Convention* (2018)
63. Pollard, K.A., Oiknine, A.H., Files, B.T., Sinatra, A.M., Patton, D., Ericson, M., Thomas, J., Khooshabeh, P.: Level of immersion affects spatial learning in virtual environments: results of a three-condition within-subjects study with long intersession intervals. *Virtual Reality* pp. 1–14 (2020)
64. Postma, B.N.J., Katz, B.: Creation and calibration method of acoustical models for historic virtual reality auralizations. *Virtual Reality* **19**, 161–180 (2015)
65. Remaggi, L., Jackson, P.J.B., Coleman, P.: Estimation of room reflection parameters for a reverberant spatial audio object. In: *Proc. of the 138th AES Convention* (2015)
66. Remaggi, L., Kim, H., Neidhardt, A., Hilton, A., Jackson, P.J.: Perceived quality and spatial impression of room reverberation in vr reproduction from measured images and acoustics. In: *Proc. ICA* (2019)
67. Ricoh: Ricoh theta v. <https://theta360.com/en/about/theta/v.html> (2019)
68. Robotham, T., Rummukainen, O., Herre, J., Habets, E.A.P.: Online vs offline multiple stimulus audio quality evaluation for virtual reality. In: *Proc. of the 145th AES Convention* (New York, USA, 2018)
69. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional networks for biomedical image segmentation. In: N. Navab, J. Hornegger, W.M. Wells, A.F. Frangi (eds.) *Proc. MICCAI*, pp. 234–241 (2015)
70. Rossing, T.D.: *Springer Handbook of Acoustics - 2nd Ed.* Springer-Verlag Berlin Heidelberg (2014)
71. Rossiter, D., Baci, G., Horner, A.: An investigation into the modelling of virtual objects with sound vibration properties. *Virtual Reality* **1**, 117–121 (1995)
72. Rothman, K.J.: No adjustments are needed for multiple comparisons. *Epidemiology* **1**, 43–46 (1990)
73. Ruminski, D.: An experimental study of spatial sound usefulness in searching and navigating through ar environments. *Virtual Reality* **19**, 223–233 (2015)
74. Schissler, C., Loftin, C., Manocha, D.: Acoustic classification and optimization for multi-modal rendering of real-world scenes. *IEEE Transactions on Visualization and Computer Graphics* **24**(3), 1246–1259 (2018)
75. Silberman, N., Hoiem, D., Kohli, P., Fergus, R.: Indoor segmentation and support inference from RGBD images. In: A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, C. Schmid (eds.) *Proc. ECCV*, pp. 746–760 (2012)
76. Smith, L.N.: A disciplined approach to neural network hyperparameters: Part 1 - learning rate, batch size, momentum, and weight decay. *CoRR abs/1803.09820* (2018)
77. Song, M., Watanabe, H., Hara, J.: Robust 3d reconstruction with omni-directional camera based on structure from motion. In: *Proc. IWAIT*, pp. 1–4 (2018)
78. Song, S., Yu, F., Zeng, A., Chang, A.X., Savva, M., Funkhouser, T.: Semantic Scene Completion from a Single Depth Image. In: *Proc. CVPR* (2017)
79. Song, S., Zeng, A., Chang, A.X., Savva, M., Savarese, S., Funkhouser, T.: Im2Pano3D: Extrapolating 360° structure and semantics beyond the field of view. In: *Proc. CVPR* (2018)
80. Stan, G.B., Embrechts, J.J., Archambeau, D.: Comparison of different impulse response measurement techniques. *J. of the Audio Engineering Society* **50**(4), 249–262 (2002)
81. Sthle, L., Wold, S.: Analysis of variance (anova). *Chemometrics and Intelligent Laboratory Systems* **6**(4), 259 – 272 (1989)
82. Student: The probable error of a mean. *Biometika* **6**, 1–25 (1908)
83. Tervo, S., Patynen, J., Kuusinen, A., Lokki, T.: Spatial decomposition method for room impulse responses. *The Journal of the Audio Engineering Society* **61**(1/2), 17–28 (2013)
84. Turk, M.: Multimodal interaction: A review. *Pattern Recognition Letters* **36**, 189–195 (2014)
85. Unity: <https://unity.com/> (2019)
86. Valimaki, V., Parker, J., Savioja, L., Smith, J., Abel, J.: Fifty years of artificial reverberation. *IEEE Transactions on Audio, Speech, and Language Processing* **20**(5), 1421–1448 (2012)
87. Vorländer, M.: International round robin on room acoustical computer simulations. In: *Proc. of the 15th ICA* (Trondheim, Norway, 1995)
88. Zhang, J., Zhao, H., Yao, A., Chen, Y., Zhang, L., Liao, H.: Efficient semantic scene completion network with spatial group convolution. In: *Proc. ECCV* (2018)