Virtual Reality
https://doi.org/10.1007/s10055-021-00594-3

ORIGINAL ARTICLE

Immersive audio-visual scene reproduction using semantic scene reconstruction from 360 cameras

Hansung Kim[1] · Luca Remaggi[2] · Aloisio Dourado[3] · Teofilo de Campos[3] · Philip J. B. Jackson[4] · Adrian Hilton[4]

Received: 9 October 2020 / Accepted: 4 October 2021

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 omnidirectional 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 adapt 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

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 (Hicks et al. 2004; Larsson et al. 2010). 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 (Turk 2014; Ruminski 2015).

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 (Mekuria et al. 2017), education (Bhama et al. 2017; Pollard et al. 2020), health care (Laver et al. 2015), entertainment (Narayanan et al. 2020) and media production (Cosker et al. 2013; Kim et al. 2012). 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 (Gonzalez-Franco and Lanier 2017). For example, sounds should be provided with the correct early reflections and reverberation effect which the user expects from the visual scene (Bailey and Fazenda 2017). This also allows correct perception of distance to the sound source (Neidhardt et al. 2018). Many studies performed evaluations of reconstructed 3D visual scenes (Menzies et al. 2016; Kim et al. 2020), but the quality of the audio has not been considered. Some researches investigated virtual reality auralisations (Rossiter et al. 1995; Postma and Katz 2015) 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

✉ Hansung Kim
h.kim@soton.ac.uk

Luca Remaggi
luca_remaggi@cle.creative.com

Aloisio Dourado
aloisio.dourado.bh@gmail.com

Teofilo de Campos
t.decampos@oxfordalumni.org

Philip J. B. Jackson
p.jackson@surrey.ac.uk

Adrian Hilton
a.hilton@surrey.ac.uk

[1] ECS, University of Southampton, Southampton, UK

[2] Creative Labs UK, London, UK

[3] University of Brasilia, Brasilia, Brazil

[4] CVSSP, University of Surrey, Guildford, UK

Published online: 30 October 2021

microphone/speaker pair recording (Li et al. 2018) and self-supervised deep learning (Pedro Morgado Nuno Vasconcelos and Wang 2018).

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 omnidirectional (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 (Kim et al. 2019), 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 (Kim et al. 2019) 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) (Furukawa and Hernández 2015), simultaneous localisation and mapping (SLAM) (Cadena et al. 2016) and structure from motion (SfM) (Bianco et al. 2018) 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 (Newcombe et al. 2011). 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 (Peng et al. 2015; Barazzetti et al. 2018) can provide a good solution for this coverage problem. Song et al. proposed a SfM method from a 360° camera (Song et al. 2018). Im et al. proposed a dense depth map estimation pipeline using a narrow-baseline video clip captured by a 360° camera (Im et al. 2016). We also proposed scene reconstruction methods using stereo 360° images from various types of 360° cameras (Kim and Hilton 2013; Kim et al. 2019). 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. (2017). 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. (2018) proposed to use Spatial Group Convolutions to reduce the amount of computational resources for network training. Liu et al. (2018) 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. (2020) proposed a 3-D scene graph for a semantic representation of rooms. We proposed EdgeNet (Dourado et al. 2021), 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 (Armeni et al. 2016) and Matterport 3D (Chang et al. 2017), 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 eight 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 (Valimaki et al. 2012; Remaggi et al. 2015; Politis et al. 2018). The parameters are typically extracted from measured acoustical room impulse responses (RIRs) (Tervo et al. 2013; Politis et al. 2018). For VR scenes, RIRs can be synthesised from the room geometry (Kim et al. 2017). 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 (Kon and Koike 2018). 3D models with material information allow the emulation of real world acoustics (Hulusic et al. 2012). Li et al. proposed scene-aware spatial audio reproduction from a single video recording (Li et al. 2018) 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 (Schissler et al. 2018), 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 (Gaudio: Gaudio vr audio 2021). 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 (Kinetic 2021). Google Resonance (Google: Google resonance audio 2021) and steam audio (Corporation 2021) 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 Sect. 3.4.

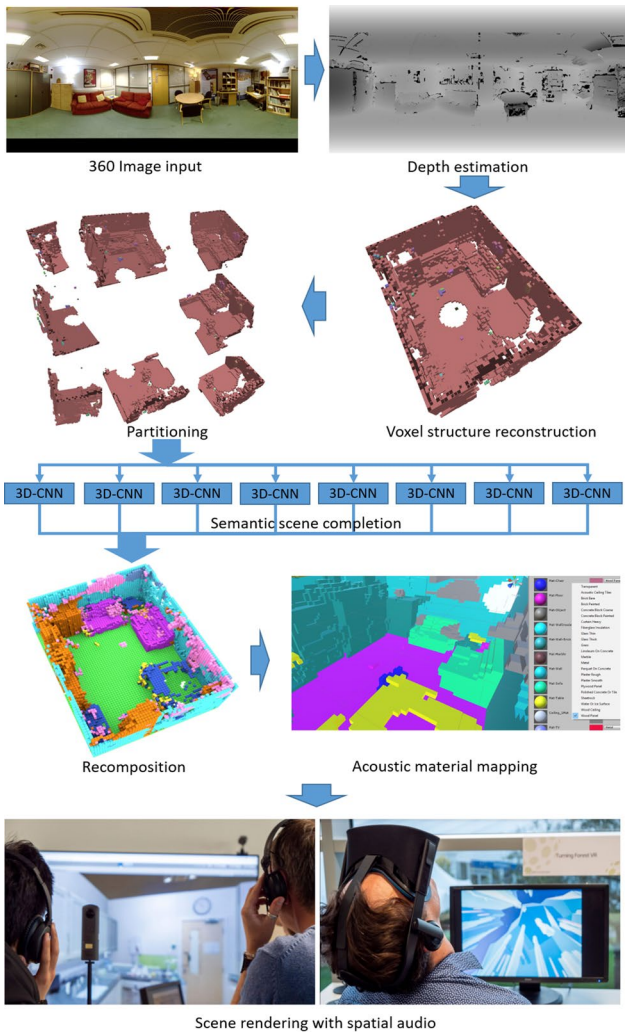


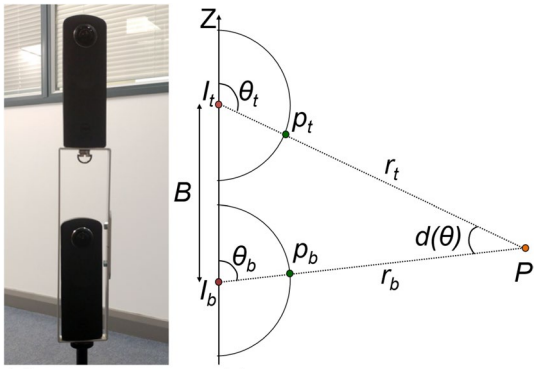
Fig. 1 Overview of the proposed pipeline

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



(a) Camera set up **(b)** Spherical stereo geometry

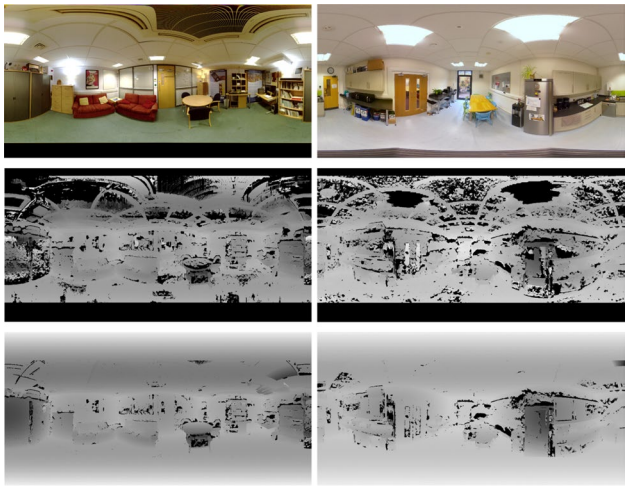
Fig. 2 Spherical stereo geometry with 360° cameras

into eight overlapped parts. The partitioned voxel structures are 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 (Unity 2019) 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 (Insta360: Insta360 one × 2019; GoPro: Gopro fusion 2019). Ricoh Theta cameras (Ricoh: Ricoh theta v 2019) were used in our system as they provide accurately rectified equi-rectangular photographs 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 to extract depth information as shown in Fig. 2a. From the pair of vertical stereo 360° images, depth of the scene is estimated by dense correspondence matching (Kim and Hilton 2013). According to the spherical stereo geometry in Fig. 2b, 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:



(a) Meeting Room (MR) **(b)** Kitchen (KT)

Fig. 3 Depth enhancement (Top: original top image; Middle: estimated depth map; Bottom: enhanced depth map)

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

Any correspondence matching algorithm can be used for the proposed pipeline. We used a simple feature-based bidirectional 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 (Gupta et al. 2010) and edge information is a distinguishing feature for reliable stereo matching, providing good depth estimates on their neighbourhood (Bleyer and Breiteneder 2013). The Canny edge detector (Gonzalez and Woods 2017) 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-coloured walls and table tops. Planes are fitted to those regions using RANSAC (Fischler and Bolles 1981) 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

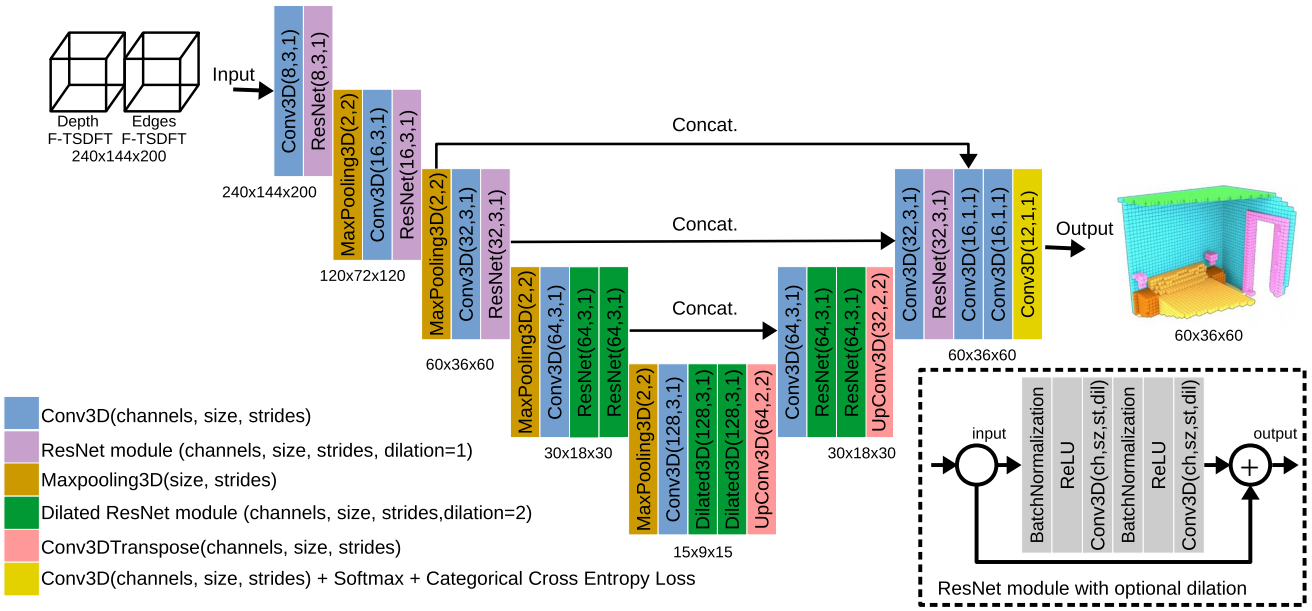


Fig. 4 U-shaped architecture for complete 3D semantic scene reconstruction

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 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 (Dourado et al. 2021). 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 eight partially overlapped views from the centre 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 1.7 m 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 0.02 m 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 (Song et al. 2017) and fine-tuned on the NYU depth v2 (NYU-v2) training set (Silberman et al. 2012). 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 voxelising the 3D mesh annotations from Guo et al. (2015) and mapped object categories based on Handa et al. (2015) 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 airtight 3D structure with object labels from depth and edge maps. This has been inspired by the U-Net design (Ronneberger et al. 2015), and differs from other Semantic Scene Completion approaches by fusing both depth and edges after encoding using F-TSDF (Song et al. 2018). 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 (Smith 2018) combined with curriculum learning (Bengio et al. 2009) and simulated annealing (Aarts and Korst 1989) 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) optimiser 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 eight object-labelled 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. (1998) 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:

$$S_i = \sum_{j=1}^{n} P_{ij} \,. \tag{2}$$

The winning class C for this voxel is:

$$C = \arg\max_i (S_i) \,. \tag{3}$$

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 (Google: Google resonance audio 2021) and steam audio (Corporation 2021). They both implement their spatial auralisation by employing binaural RIRs (BRIRs) over virtual loudspeakers.

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

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.[1]

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 (Robotham et al. 2018; Gorzel et al. 2019). 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 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 spatialisation.[2] Therefore, we used Steam Audio for our final implementation though we still included the results with

[1] https://valvesoftware.github.io/steam-audio/doc/phonon_unity.html.

[2] https://resonance-audio.github.io/resonance-audio/develop/design-tips.html.

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.

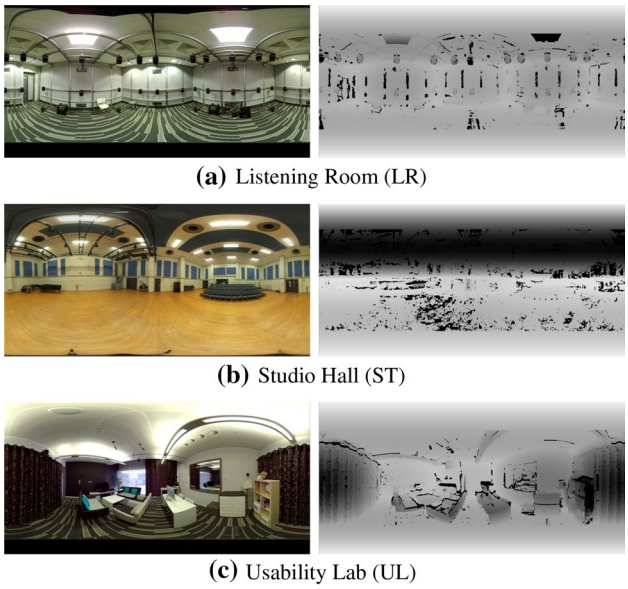
Although measuring RIRs in real environments is well-established (Stan et al. 2002), extracting RIR information from VR environments has not previously been explored. Therefore, we treat the virtual environment as a real one to measure BRIRs and emulate virtual binaural microphones and omnidirectional sound sources in the reconstructed virtual environments, to record sounds. The general swept-sine method (Farina 2000) is employed to calculate RIRs for Steam Audio, and an anechoic gun-shot (normalised in the time domain) (Cox 2013) 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.

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 (HTC: Vive pro 2018), 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 "plausibity" of rendered sound should be considered. Authenticity measures how identical the generated sound is to the ground truth sound (Blauert 2005), while plausibility is subjectively judged by the listener with his/her inner reference (listener's expectation) (Lindau and Weinzierl 2012). 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 evaluated 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.

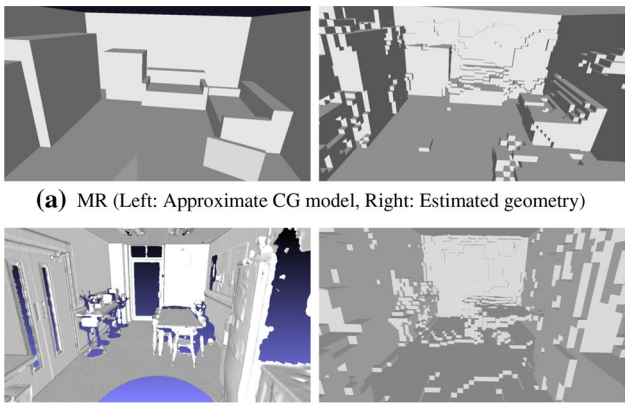


(a) Listening Room (LR)

(b) Studio Hall (ST)

(c) Usability Lab (UL)

Fig. 5 Dataset used in the experiments (Left: captured top image; Right: enhanced depth map)



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

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

Fig. 6 Comparison of geometry reconstruction

All datasets, 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

Table 2 Comparison of reconstruction errors in room dimensions

Data	Ground truth	Kim19 (Kim et al. 2019)		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, **1.29**)	5.54 × 4.24 × 2.40	(**1.25**, **0.93**, 3.00)
KT	6.64 × 3.46 × 2.67	6.95 × 3.41 × 2.70	(4.67, **1.45**, 1.12)	6.42 × 3.52 × 2.68	(**3.31**, 1.73, **0.37**)
LR	5.64 × 5.05 × 2.90	5.77 × 5.17 × 2.98	(**2.30**, 2.38, **2.76**)	5.88 × 5.02 × 2.78	(4.26, **0.59**, 4.14)
ST	17.08 × 14.55 × 6.50	16.53 × 14.87 × 5.70	(3.22, **2.20**, **12.31**)	17.54 × 15.46 × 5.56	(**2.69**, 6.25, 14.46)
UL	5.57 × 5.20 × 2.91	5.92 × 4.95 × 2.95	(6.28, 4.81, **1.37**)	5.52 × 5.22 × 3.00	(**0.90**, **0.38**, 3.09)

The bold figures represent the minimum error in each dimension

(Kim19) (Kim et al. 2019) which employs two separate processes of 2D semantic segmentation by SegNet (Badrinarayanan et al. 2017) 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 Kim et al. (2019) and the proposed pipeline, respectively. In the outputs of Kim et al. (2019), 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 Kim et al. (2019), 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 Kim et al. (2019) and the proposed method built relatively accurate room layouts. The proposed method shows slightly better estimations for room width and length, but Kim et al. (2019) were a bit better in the room height estimation because the depth for the whole ceilings and floors were inferred from the 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 inversely 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. Kim et al. (2019) 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 (Kim et al. 2019) which requires two separate processes for 8 mins (3 mins of 2D semantic segmentation and 5 mins of 3D geometry reconstruction) (Fig. 7).

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 (Farina 2000) 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 spatial effects that are encoded in timbral attributes 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.

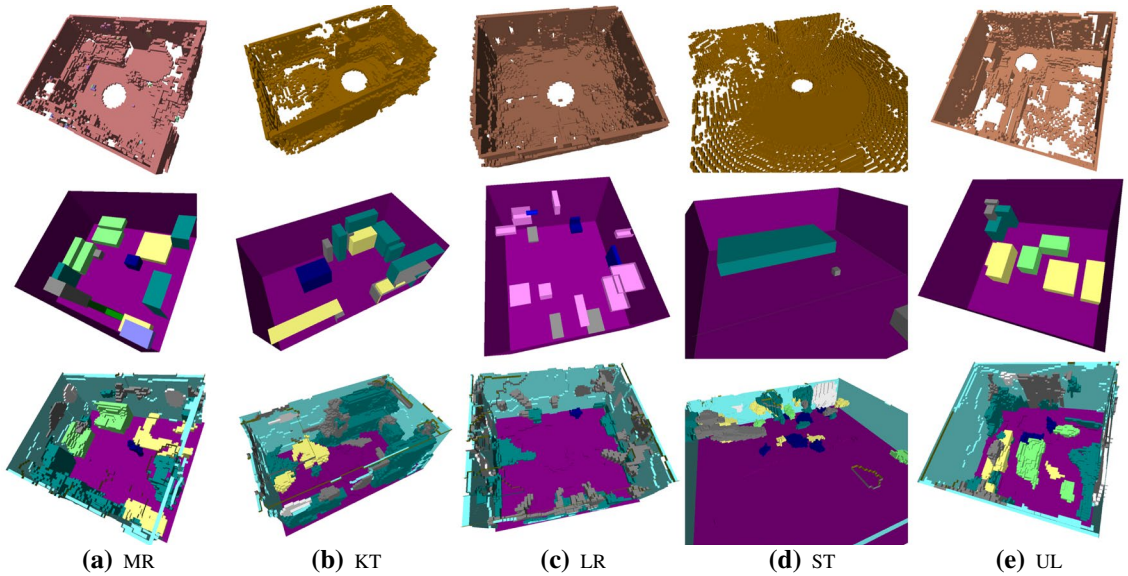


Fig. 7 Semantic room reconstruction results (Top: initial voxel cloud, Middle: reconstruction by Kim et al. (2019), Bottom: proposed method, the colour index is defined as in Fig. 8

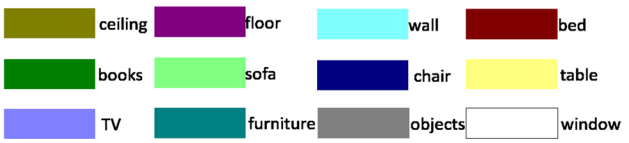


Fig. 8 Colour index for semantic objects linked to the materials in Table 1

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 (Bradley 2011; Rossing 2014). On the other hand, RT60 relates to the average absorption, location of room boundaries and size of the room, describing the reverberation from a physical point of view (Bradley 2011; Rossing 2014). EDT is calculated as six times the time required for the energy to decay 10 dB after the direct sound (Barron 1995). RT60 is measured as the time for the energy to decay 60 dB. The average values over the six 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 (Meng et al. 2006) and 5% for EDT (Vorländer 1995) as suggested in the literature. However, it is important to note that the same literature describes variable percentage depending on the type of sounds. For instance, in Meng et al. (2006), 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 (Lindau and Weinzierl 2012; Blauert 2005). 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 (Neidhardt et al. 2018). 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 (Bailey and Fazenda 2018). 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 Figs. 9 and 10, respectively. The proposed method was compared along with three other models. First, "Kim19" is a state-of-the-art method (Kim et al. 2019)

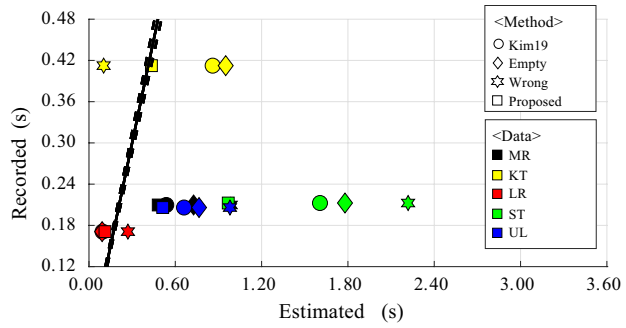


Fig. 9 EDTs for the five rooms, related to the estimated RIRs in VR environment. The dashed lines show the JND limit of 5% (Vorländer 1995)

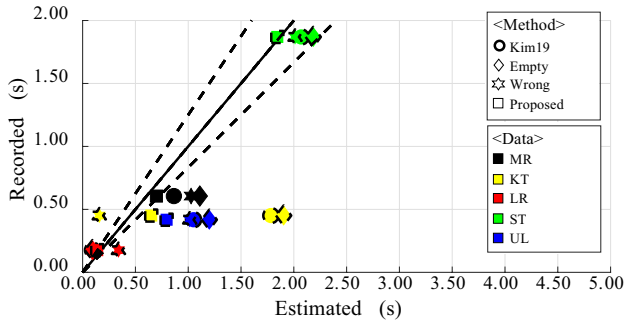


Fig. 10 RT60s for the five rooms, related to the estimated RIRs in VR environment. The dashed lines show the JND limit of 20% (Meng et al. 2006)

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, 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 acoustics. "Empty" and "Wrong" give the worst performances 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 (Kim et al. 2019)" 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 wrong materials comparing with the "Kim19 (Kim et al. 2019)" models which have very simple box-shaped geometry.

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

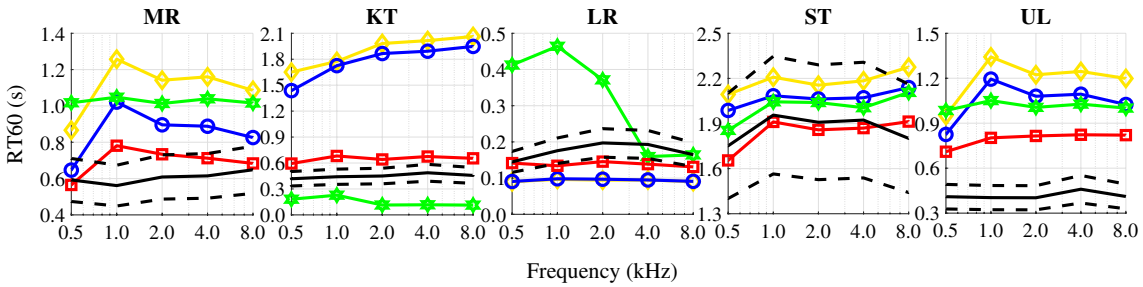


Fig. 11 RT60s over the different frequency bands for the five rooms. (Black line: ground truth, Dashed black lines: JND, Blue circle: Kim19 (Kim et al. 2019). Yellow diamond: Empty, Green star: Wrong, Red square: Proposed)

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[3], 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 (Garofolo et al. 1993) and a clarinet sound in an anechoic chamber from the OpenAirLib library (Brown et al. 2017).

In order to evaluate subjective attributes of the reproduced sound quality, two factors were tested as proposed in Hoeg et al. (1997): "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

[3] https://github.com/IoSR-Surrey/MUSHRA-MaxMSP.

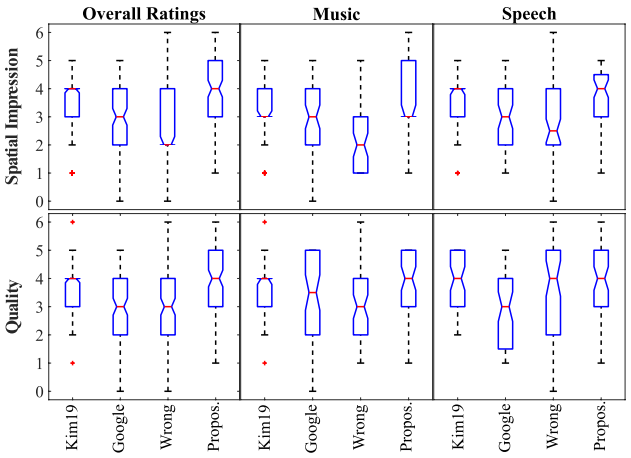


Fig. 12 Subjective scores for spatial impression (top three) and overall quality (bottom three). On each box, the horizontal red line represents the median of the distribution, the bottom and top edges of the boxes are the 25th and 75th percentiles, respectively. The whiskers show the most extreme non-outlier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

(Kim et al. 2019) used in previous Sects. 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 Sect. 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.

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" (Kim et al. 2019), but looking at their 75th percentile, the results related to the proposed method are more stretched towards values greater than 4, as opposed to the results related to "Kim19" (Kim et al. 2019). It is also important to show that participants clearly perceived the lowest spatial impression when the

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

	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

$H = 1$ means that the test rejects the null hypothesis of the groups belonging to normal distributions with equal means (95% of confidence), also denoted by *

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

	Overall ratings				Music				Speech			
	Kim19	Google	Wrong	Prop.	Kim19	Google	Wrong	Prop.	Kim19	Google	Wrong	Prop.
SI-means	3.41	2.97	2.57	**3.63**	3.25	2.95	2.30	**3.54**	3.57	3.00	2.84	**3.73**
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	**3.72**	3.54	3.34	3.07	**3.64**	**3.88**	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	–	–	–	$H = 1$	–	–	–	$H = 1$	–	–	–	$H = 0$

$H = 1$ 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

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 75th percentile.

These plots could, mistakenly, lead to the conclusion that the four tested methods produce the 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 (Sthle and Wold 1989), 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 (Student: The probable error of a mean 1908). 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.

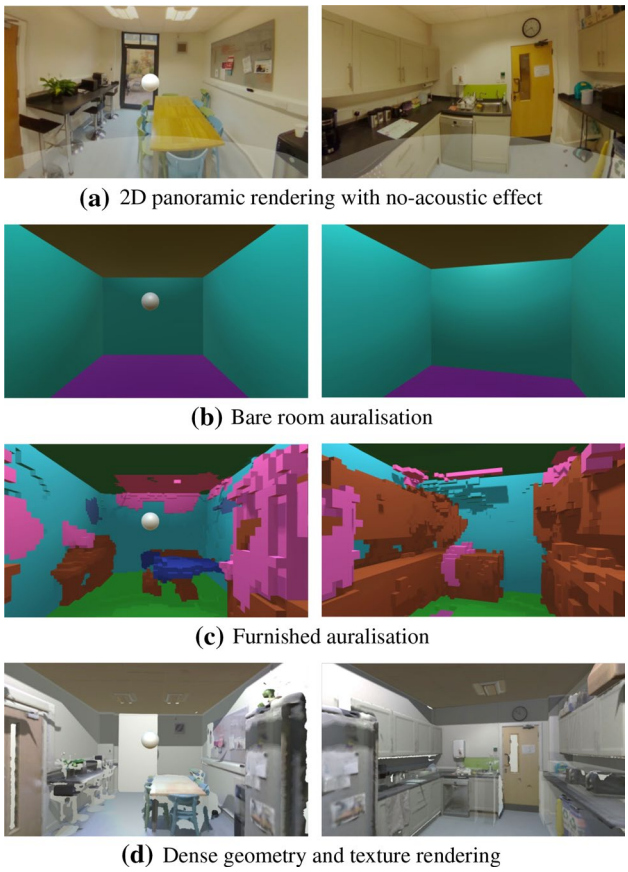
The results are reported in the bottom row of Fig. 12. The observed trend is similar to the one discussed for the spatial impression. In general, the participants seem to prefer the quality provided by the proposed method. Nonetheless, Kim19 (Kim et al. 2019) also gives good performance, with the median being the same as the new method but its 75th 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 Sect. 3.4: 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 third 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 delay to the centre 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 (Remaggi et al. 2019). 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 (Sthle and Wold 1989; Rothman 1990), 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 the 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.



(a) 2D panoramic rendering with no-acoustic effect

(b) Bare room auralisation

(c) Furnished auralisation

(d) Dense geometry and texture rendering

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

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 (HTC: Vive pro 2018) 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. 13a 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); and (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. 13a, the 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 (Bailey and Fazenda 2018). 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.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10055-021-00594-3.

Acknowledgements This work was supported by the UKRI EPSRC Programme Grant S3A: Future Spatial Audio for an Immersive Listener Experience at Home (EP/L000539/1) and Prosperity Partnership AI4ME: AI for Personalised Media Experiences EP/V038087/1, 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

Aarts E, Korst J (1989) Simulated annealing and Boltzmann machines: a stochastic approach to combinatorial optimization and neural computing. Wiley, New York

Armeni I, Sener O, Zamir AR, Jiang H, Brilakis I, Fischer M, Savarese S (2016) 3D semantic parsing of large-scale indoor spaces. In: Proceedings of CVPR, pp 1534–1543

Badrinarayanan V, Kendall A, Cipolla R (2017) Segnet: a deep convolutional encoder-decoder architecture for image segmentation. IEEE Trans Pattern Anal Mach Intell

Bailey W, Fazenda BM (2017) The effect of reverberation and audio spatialization on egocentric distance estimation of objects in stereoscopic virtual reality. J. Acoust Soc Am 141(5):3510

Bailey W, Fazenda BM (2018) The effect of visual cues and binaural rendering method on plausibility in virtual environments. In: Proceedings of the 144th AES convention, Milan, Italy

Barazzetti L, Previtali M, Roncoroni F (2018) Can we use low-cost 360 degree cameras to create accurate 3d models? ISPRS Int Arch Photogram Remote Sens Spat Inf Sci XLII–2:69–75

Barron M (1995) Interpretation of early decay times in concert auditoria. Acta Acustica 81(4):320–331

Bengio Y, Louradour J, Collobert R, Weston, J (2009) Curriculum learning. In: Proceedings of ICML, pp 41–48

Bhama PRKS, Hariharasubramanian V, Mythili OP, Ramachandran M (2017) Users-domain knowledge prediction in e-learning with speech-interfaced augmented and virtual reality contents. Virt Real 24:163–173

Bianco S, Ciocca G, Marelli D (2018) Evaluating the performance of structure from motion pipelines. J Imaging 4(8):98

Blauert J (2005) Communication acoustics. Springer, Berlin

Bleyer M, Breiteneder C (2013) Stereo matching-state-of-the-art and research challenges. In: Advanced topics in computer vision, pp 143–179

Bradley JS (2011) Review of objective room acoustics measures and future needs. Appl Acoust 72(10):713–720

Brown K, Paradis M, Murphy D (2017) Openairlib: a javascript library for the acoustics of spaces. In: Audio engineering society convention, p 142

Cadena C, Carlone L, Carrillo H, Latif Y, Scaramuzza D, Neira J, Reid I, Leonard J (2016) Past, present, and future of simultaneous

localization and mapping: towards the robust-perception age. IEEE Trans Rob 32(6):1309–1332
Chang A, Dai A, Funkhouser T, Halber M, Niessner M, Savva M, Song S, Zeng A, Zhang Y (2017) Matterport 3D: learning from RGB-D data in indoor environments. In: Proceedings of 3DV
Corporation V (2021) Steam audio. https://valvesoftware.github.io/steam-audio/
Cosker D, Eisert P, Grau O, Hancock PJB, McKinnell J, Ong E (2013) Applications of face analysis and modeling in media production. IEEE Multimed 20(4):18–27
Cox T (2013) Gun shot in anechoic chamber. Freesound. https://freesound.org/people/acs272/sounds/210766/
Dourado A, de Campos TE, Kim H, Hilton A (2021) EdgeNet: semantic scene completion from rgb-d images. In: Proceedings of ICPR
Farina A (2000) Simultaneous measurement of impulse response and distortion with a swept-sine technique. In: Proceedings of the AES convention
Fischler MA, Bolles RC (1981) Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Commun ACM 24(6):381–395
Furukawa Y, Hernández C (2015) Multi-view stereo: a tutorial. Found Trends Comput Gr Vis 9(1–2):1–148
Garofolo JS, Lamel LF, Fisher WM, Fiscus JG, Pallet DS, Dahlgren NL (1993) DARPA TIMIT acoustic phonetic continuous speech corpus CDROM. Technical report. NIST Interagency
Gaudio: Gaudio vr audio (2021). https://gaudiolab.com/solution-ar-vr-and-immersive/
Gonzalez R, Woods R (2017) Digital image processing. Pearson
Gonzalez-Franco M, Lanier J (2017) Model of illusions and virtual reality. Front Psychol 8(1):1125
Google: Google resonance audio (2021). https://resonance-audio.github.io/resonance-audio/
GoPro: Gopro fusion (2019). https://shop.gopro.com/EMEA/cameras/fusion/CHDHZ-103-master.html
Gorzel M, Allen A, Kelly I, Gungormusler A, Kammerl J, Yeh H, Boland F (2019) Efficient encoding and decoding of binaural sound with resonance audio. In: Proceedings of the AES conference on immersive and interactive audio, York, UK
Guo R, Zou C, Hoiem D (2015) Predicting complete 3D models of indoor scenes. CoRR abs/1504.02437. http://arxiv.org/abs/1504.02437
Gupta A, Efros AA, Hebert M (2010) Blocks world revisited: image understanding using qualitative geometry and mechanics. In: Proceedings of ECCV
Handa A, Patraucean V, Badrinarayanan V, Stent S, Cipolla R (2015) SceneNet: understanding real world indoor scenes with synthetic data. CoRR abs/1511.07041. http://arxiv.org/abs/1511.07041
Hicks M, Nichols S, O'Malley C (2004) Comparing the roles of 3d representations in audio and audio-visual collaborations. Virt Real 7:148–163
Hoeg W, Christensen L, Walker R (1997) Subjective assessment of audio quality—the means and methods within the EBU. Technical report. EBU Technical Review
HTC: Vive pro (2018). https://www.vive.com/uk/product/vive-pro-full-kit/
Hulusic V, Harvey C, Debattista K, Tsingos N, Walker S, Howard D, Chalmers A (2012) Acoustic rendering and auditory-visual cross-modal perception and interaction. J Comput Gr Forum 31(1):102–131
Im S, Ha H, Rameau F, Jeon HG, Choe G, Kweon I (2016) All-around depth from small motion with a spherical panoramic camera. In: Proceedings of ECCV, vol 9907
Insta360: Insta360 one x (2019). https://www.insta360.com/product/insta360-onex
Kim H, Guillemaut JY, Takai T, Sarim M, Hilton A (2012) Outdoor dynamic 3d scene reconstruction. IEEE Trans Circuits Syst Video Technol 22(11):1611–1622
Kim H, Hilton A (2013) 3D scene reconstruction from multiple spherical stereo pairs. Int J Comput Vis 104(1):94–116
Kim H, Hughes RJ, Remaggi L, Jackson PJB, Hilton A, Cox TJ, Shirley B (2017) Acoustic room modelling using a spherical camera for reverberant spatial audio objects. In: Proceedings of the 142th AES convention
Kim H, Remaggi L, Jackson PJ, Hilton A (2019) Immersive spatial audio reproduction for vr/ar using room acoustic modelling from 360 images. In: Proceedings of IEEE VR
Kim HG, Lim H, Ro YM (2020) Deep virtual reality image quality assessment with human perception guider for omnidirectional image. IEEE Trans Circuits Syst Video Technol 30(4):917–928
Kim U, Park J, Song T, Kim J (2020) 3-d scene graph: a sparse and semantic representation of physical environments for intelligent agents. IEEE Trans Cybern 50(12):4921–4933
Kinetic A (2021) Wwise spatial audio. https://www.audiokinetic.com/products/wwise-spatial-audio/
Kittler J, Hatef M, Duin RPW, Matas J (1998) On combining classifiers. IEEE Trans Pattern Anal Mach Intell 20(3):226–239
Kon H, Koike H (2018) Deep neural networks for cross-modal estimations of acoustic reverberation characteristics from two-dimensional images. In: Proceedings of the 144th AES convention, Milan, Italy
Larsson P, Väljamäe A, Västfjäll D, Tajadura-Jiménez A, Kleiner M (2010) Auditory-induced presence in mixed reality environments and related technology
Laver KE, George S, Thomas S, Deutsch JE, Crotty M (2015) Virtual reality for stroke rehabilitation. Cochrane Collab 2:1–27
Li D, Langlois TR, Zheng C (2018) Scene-aware audio for 360° videos. ACM Trans Gr 37(4)
Lindau A, Weinzierl S (2012) Assessing the plausibility of virtual acoustic environments. Acta Acust Acust 98(5):804–810
Liu S, Hu Y, Zeng Y, Tang Q, Jin B, Han Y, Li X (2018) See and think: disentangling semantic scene completion. In: Bengio S, Wallach H, Larochelle H, Grauman K, Cesa-Bianchi N, Garnett R (eds) Proceedings of NIPS, pp 263–274
Mekuria R, Blom K, Cesar P (2017) Design, implementation, and evaluation of a point cloud codec for tele-immersive video. IEEE Trans Circuits Syst Video Technol 27(4):828–842
Meng Z, Zhao F, He M (2006) The just noticeable difference of noise length and reverberation perception. In: Proceedings of the international symposium on communications and information technologies, Bangkok, Thailand
Menzies R, Rogers SJ, Phillips AM, Chiarovano E, Waele C, Verstraten F, MacDougall H (2016) An objective measure for the visual fidelity of virtual reality and the risks of falls in a virtual environment. Virt Real 20:173–181
Narayanan S, Polys N, Bukvic I (2020) Cinemacraft: exploring fidelity cues in collaborative virtual world interactions. Virt Real 24:53–73
Neidhardt A, Tommy AI, Pereppadan AD (2018) Plausibility of an interactive approaching motion towards a virtual sound source based on simplified BRIR sets. In: Proceedings of the 144th AES convention, Milan, Italy
Newcombe R, Izadi S, Hilliges O, Molyneaux D, Kim D, Davison A, Kohli P, Shotton J, Hodges S, Fitzgibbon A (2011) Kinectfusion: real-time dense surface mapping and tracking. In: Proceedings of ISMAR
Morgado P, Vasconcelos N, Langlois T, Wang O (2018) Self-supervised generation of spatial audio for 360° video. In: Proceedings of NIPS

Peng X, Bennamoun M, Wang Q, Ma Q, Xu Z (2015) A low-cost implementation of a 360° vision distributed aperture system. IEEE Trans Circuits Syst Video Technol 25(2):225–238

Politis A, Tervo S, Lokki T, Pulkki V (2018) Parametric multidirectional decomposition of microphone recordings for broadband high-order ambisonic encoding. In: Proceedings of the 144th AES convention

Pollard KA, Oiknine AH, Files BT, Sinatra AM, Patton D, Ericson M, Thomas J, Khooshabeh P (2020) Level of immersion affects spatial learning in virtual environments: results of a three-condition within-subjects study with long intersession intervals. Virt Real 1–14

Postma BNJ, Katz B (2015) Creation and calibration method of acoustical models for historic virtual reality auralizations. Virt Real 19:161–180

Remaggi L, Jackson PJB, Coleman P (2015) Estimation of room reflection parameters for a reverberant spatial audio object. In: Proceedings of the 138th AES convention

Remaggi L, Kim H, Neidhardt A, Hilton A, Jackson PJ (2019) Perceived quality and spatial impression of room reverberation in VR reproduction from measured images and acoustics. In: Proceedings of ICA

Ricoh: Ricoh theta v (2019). https://theta360.com/en/about/theta/v.html

Robotham T, Rummukainen O, Herre J, Habets EAP (2018) Online vs offline multiple stimulus audio quality evaluation for virtual reality. In: Proceedings of the 145th AES convention, New York, USA

Ronneberger O, Fischer P, Brox,T (2015) U-Net: convolutional networks for biomedical image segmentation. In: Navab N, Hornegger J, Wells WM, Frangi AF (eds) Proceedings of MICCAI, pp 234–241

Rossing TD (2014) Springer handbook of acoustics, 2nd edn. Springer, Berlin

Rossiter D, Baciu G, Horner A (1995) An investigation into the modelling of virtual objects with sound vibration properties. Virt Real 1:117–121

Rothman KJ (1990) No adjustments are needed for multiple comparisons. Epidemiology 1:43–46

Ruminski D (2015) An experimental study of spatial sound usefulness in searching and navigating through AR environments. Virt Real 19:223–233

Schissler C, Loftin C, Manocha D (2018) Acoustic classification and optimization for multi-modal rendering of real-world scenes. IEEE Trans Visual Comput Gr 24(3):1246–1259

Silberman N, Hoiem D, Kohli P, Fergus R (2012) Indoor segmentation and support inference from RGBD images. In: Fitzgibbon A, Lazebnik S, Perona P, Sato Y, Schmid C (eds) Proceedings of ECCV, pp 746–760

Smith LN (2018) A disciplined approach to neural network hyperparameters: part 1—learning rate, batch size, momentum, and weight decay. CoRR abs/1803.09820

Song M, Watanabe H, Hara J (2018) Robust 3d reconstruction with omni-directional camera based on structure from motion. In: Proceedings of IWAIT, pp 1–4

Song S, Yu F, Zeng A, Chang AX, Savva M, Funkhouser T (2017) Semantic scene completion from a single depth image. In: Proceedings of CVPR

Song S, Zeng A, Chang AX, Savva M, Savarese S, Funkhouser T (2018) Im2Pano3D: extrapolating 360° structure and semantics beyond the field of view. In: Proceedings of CVPR

Stan GB, Embrechts JJ, Archambeau D (2002) Comparison of different impulse response measurement techniques. J Audio Eng Soc 50(4):249–262

Sthle L, Wold S (1989) Analysis of variance (anova). Chemom Intell Lab Syst 6(4):259–272

Student (1908) The probable error of a mean. Biometika 6:1–25

Tervo S, Patynen J, Kuusinen A, Lokki T (2013) Spatial decomposition method for room impulse responses. J Audio Eng Soc 61(1/2):17–28

Turk M (2014) Multimodal interaction: a review. Pattern Recogn Lett 36:189–195

Unity (2019). https://unity.com/

Valimaki V, Parker J, Savioja L, Smith J, Abel J (2012) Fifty years of artificial reverberation. IEEE Trans Audio Speech Lang Process 20(5):1421–1448

Vorländer M (1995) International round robin on room acoustical computer simulations. In: Proceedings of the 15th ICA, Trondheim, Norway

Zhang J, Zhao H, Yao A, Chen Y, Zhang L, Liao H (2018) Efficient semantic scene completion network with spatial group convolution. In: Proceedings of ECCV