.

3D Audio-Visual Indoor Scene Reconstruction and Semantics Completion for Virtual Reality from a Single 360° RGB-D Image

Mona Alawadh[1,2*], Atiyeh Alinaghi[1], Mahesan Niranjan[1], Hansung Kim[1*]

[1*]ECS, University of Southampton, Street, Southampton, SO17 1BJ, Hampshire, UK.

[2]CCIS, Imam Mohammad Ibn Saud Islamic University, Al Thoumamah Road, Riyadh, 11564, Riyadh, Saudi Arabia.

*Corresponding author(s). E-mail(s): malawadh@imamu.edu.sa; h.kim@soton.ac.uk;
Contributing authors: a.alinaghi@soton.ac.uk; mn@ecs.soton.ac.uk;

Abstract

We introduce a new approach for constructing immersive virtual spaces by generating comprehensive 3D voxelised models that encompass both geometric and semantic scene representations from a single 360° RGB-D input. The proposed approach utilises a deep convolutional neural network for semantic scene completion (SSC), allowing the estimation of complete semantics and geometries of the scene. We design MDBNet a dual head model that simultaneously processes RGB and depth data using a perspective camera. Depth information is encoded using a flipped transcribed signed distance function (F-TSDF), capturing essential geometric shape characteristics. We extend the inference capabilities of MDBNet on RGB-D input of the perspective camera to accommodate 360° RGB-D by proposing MDBNet360. We employ RGB spherical-to-cubic projection and 3D rotation for depth point clouds, allowing for virtual reality (VR) space design with comprehensive spatial coverage. To our knowledge, this is the first work to extend a pre-trained SSC model, originally using perspective camera RGB-D input, to infer a 3D model from 360° RGB-D input. To assess acoustic properties, we measure parameters such as early decay time (EDT) and reverberation time (RT60) using the exponential sine sweep method (ESS). We used Unity with the Steam Audio plug-in for conducting simulations in virtual space. The proposed framework demonstrates better virtual space reconstruction and immersive sound generation, advancing semantically rich and spatially accurate virtual environments compared to the state-of-the-art (SOTA). Code and rendered sounds are available on GitHub: https://github.com/MonaIA1/Repo360

Keywords: Semantic Scene Completion, 3D reconstruction, room acoustic modelling, VR

1 Introduction

In virtual reality (VR) space, humans can interact with a simulated world of three dimensions (3D) in real time, experiencing the illusion of being fully immersed in a synthetic environment [1]. Both visual and synchronised spatial audio are essential for creating truly immersive environment experiences [2–4]. The integration of both audio and visual aspects enables users to perceive a digital 3D space that closely mimics real world environments.

However, the immersion effect is based mainly on visual perception [5]. Building on the role of visual perception in immersive experiences, this research explores the application of artificial intelligence (AI) in computer vision, by utilising deep learning methods on 2D images. In our daily lives, various types of cameras, such as perspective and 360° cameras, are widely available, capturing vast amounts of 2D images, including RGB and depth maps. This research focuses on transforming 2D images into comprehensive, semantically annotated 3D models for use in VR spaces. Since 2D images capture only partial information about 3D scenes, AI enables the development of models capable of understanding and learning the underlying structure and semantics of the 3D world, including the reconstruction of occluded areas from a single 2D input, which is known as semantic scene completion (SSC). SSC aims to infer the complete 3D structure, including occluded regions, from a single perspective view [6]. SSC is a challenging and ill-posed task in computer vision, particularly for voxelised indoor environments, due to the inherently limited nature of the input and the significant loss of 3D information in unobserved areas. Furthermore, data sparsity and imbalanced class distributions in existing datasets compound the difficulty of accurate prediction. Predicting object semantics in 3D space is particularly challenging due to the complexity of inferring information about occluded or partially visible objects. Key obstacles include dataset imbalances, intraclass diversity, and interclass ambiguity [7]. The motivation for this study stems from the need to construct immersive VR spaces while simplifying the traditionally resource-intensive process of generating audio-visual scenes. Conventional methods for estimating room acoustic properties rely on physical measurements using microphones and loudspeakers, which are time-consuming and hardware-intensive [8]. In contrast, this work presents a computer vision–based approach that reconstructs the 3D geometry of a scene and estimates its acoustic parameters through synthesized room impulse responses (RIRs) derived from a single 360° RGB-D image. The proposed method offers a practical and scalable solution for dynamically creating immersive VR environments, with potential applications in entertainment, architectural design, education , and tourism.

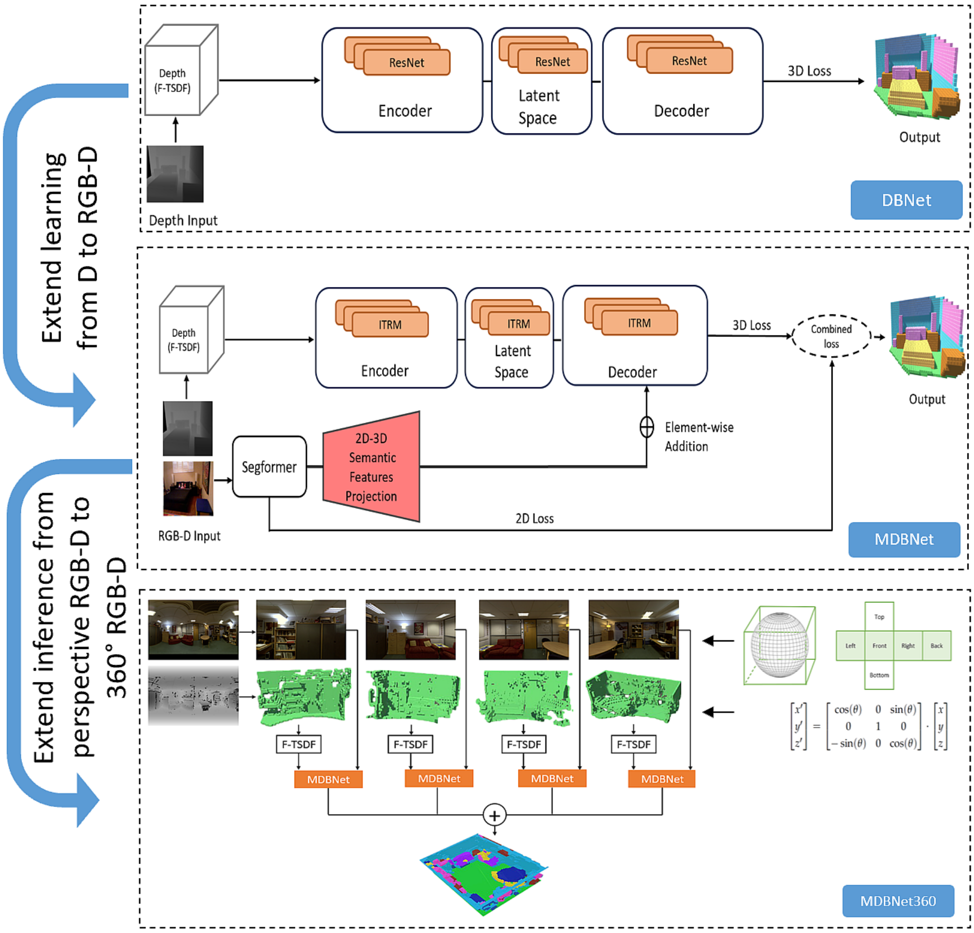

Extend learning from D to RGB-D
Extend inference from perspective RGB-D to 360° RGB-D
Depth (F-TSDF)
Depth Input
ResNet
Encoder
ResNet
Latent Space
ResNet
Decoder
3D Loss
Output
DBNet
Depth (F-TSDF)
ITRM
Encoder
ITRM
Latent Space
ITRM
Decoder
3D Loss
Combined loss
Output
Element-wise Addition
RGB-D Input
Segformer
2D-3D Semantic Features Projection
2D Loss
MDBNet
F-TSDF
MDBNet
Top
Left
Front
Right
Back
Bottom
$\begin{bmatrix} x' \\ y' \\ z' \end{bmatrix} = \begin{bmatrix} \cos(\theta) & 0 & \sin(\theta) \\ 0 & 1 & 0 \\ -\sin(\theta) & 0 & \cos(\theta) \end{bmatrix} \cdot \begin{bmatrix} x \\ y \\ z \end{bmatrix}$
MDBNet360

Fig. 1: System stages improvement to construct 3D space starting by DBNet with depth only input then MDBNet with RGB-D, ending with MDBNet360 with RGB-D input.

In this research, as illustrated in Figure 1 we extend our previous model for SSC using depth-only input [9], which we refer to as 'DBNet'. We propose a new hybrid architecture, MDBNet, which features a dual-head design that simultaneously processes RGB and depth data. Depth information is encoded using a flipped-truncated signed distance function (F-TSDF), capturing essential geometric shape characteristics. The RGB features are projected from 2D to 3D space using depth maps. We explore various RGB semantics fusion strategies, including early, middle, and late fusion methods. The proposed model enhances the performance of SSC predictions on RGB-D inputs. Furthermore, to construct VR space with 360° field-of-view (FOV), this research addresses the challenge of extending the inference of SSC from partial views to full 360° coverage, enabling the prediction of 3D annotated models from a single 2D image with full panorama. Constructing 3D models from partial views alone often falls short of providing the fully immersive experience required for realistic VR applications. To overcome this limitation, this study aims to generate complete VR spaces with 360° surroundings, creating an environment that closely mirrors the user's spatial perception in the real world. We adopt a spherical-to-cubic projection technique for RGB data and apply a 3D rotation method to depth point clouds to ensure proper alignment with the cubic projection of 2D images. Our contribution lies on demonstrating the feasibility of end-to-end 3D semantic reconstruction and completion on full panoramic scenes, where parallax information is absent (due to the single optical center of the panoramic capture) and prior SSC methods have not addressed this challenge for a single 360° RGB-D input.

To achieve enhanced immersion, spatial sound must be integrated with 3D models. In this research, sound rendering and modelling are performed using Unity [1] VR gaming engine equipped with spatial sound plug-in Steam Audio [2] to generate a 3D virtual environment of a real world space. This integration ensures an immersive auditory and visual experience, which is essential for VR applications. To measure the plausibility of the rendered sound in the VR environment, we assess the acoustic properties by measuring RIR acoustic parameters such as early decay time (EDT) and reverberation time (RT60) using the exponential sine sweep method (ESS).

This research contributes to bridge the gap between computational modelling and human perception. It introduces a horizontal integration of AI and VR, designed to support more intuitive, human-centered digital interactions. This work opens new pathways for human communication and engagement [10], and contributes to revolutionising experiential learning paradigms [11, 12]. As highlighted in recent reviews of immersive technologies and AI for human-centered digital experiences [12], such convergence blurs the boundaries between physical and digital realities, enabling adaptive, personalised, and emotionally resonant environments that reflect and expand human cognition. The summary of our contributions are as follows:

- Extend the previous work DBNet [9] and propose MDBNet SSC model with a dual-head and combined loss function to train the model simultaneously with both single RGB and depth data of perspective views. We quantify the performance uncertainty in our results to ensure an unbiased assessment across trials, contributing to more reliable benchmarking in the SSC field.
- Perform an acoustic analysis of the 3D virtual environments generated by MDBNet360 through the evaluation of RIRs acoustic parameters, such as EDT and RT60, and comparing the results with SOTA methods. The proposed method showing better 3D scene reconstruction and acoustic parameters for the virtual space compared to SOTA.
- Design VR application to demonstrate the 3D audio-visual space from single 360° RGB-D.

[1] https://unity.com/(accessed in 2025)
[2] https://github.com/ValveSoftware/steam-audio(accessed in 2025)

2 Related work

2.1 3D semantic scene completion (SSC) from single Perspective view

SSC is a relatively recent research field that began with the work by Song et al. [6], who introduced SSCNet, the first deep neural network designed specifically for SSC. The SSC task involves simultaneously predicting volumetric occupancy and object categories at the voxel level from a partial view. The design of SSC architectures is closely tied to the type of input data, including 3D geometry representations derived from depth maps using truncated signed distance function (TSDF) with volume networks, 2D inputs such as RGB and/or depth using view-volume networks, or hybrid networks that combine TSDF-based geometry representations with RGB data [13].

Several studies have utilised volume CNN designs to manage 3D scene representations through 3D occupancy grids or voxels. These grids incorporate TSDF values, typically derived from depth maps, which represent the distance to the nearest surface within a normalised range of -1 to 1 [6, 14]. Many studies, such as [6, 9, 15–18], use F-TSDF to provide steeper gradients at surface boundaries.

Other research has explored the view-volume approach, integrating 2D/3D CNNs to extract features from 2D sources like RGB and/or depth maps, and then project these features into 3D space using a projection layer [19–24]. One of the first methods to incorporate RGB features with depth data in the SSC domain was by Liu et al. [20], where the projection of 2D RGB features is based on depth maps and camera parameters. Other works, such as [25–27], utilised only RGB inputs to predict the 3D representation. However, using single RGB input alone is challenging due to the loss of depth information.

Some recent studies have shifted towards hybrid designs that utilise multiple inputs, including TSDF, RGB, or point clouds. This approach aims to leverage the strengths of both 3D geometric and 2D semantic features [14, 28–33]. SSC architectures incorporate learning from both 2D and 3D representations leveraged transfer learning to utilise the learnable feature weights from large datasets [14, 29, 31, 32, 34], with some adopting ResNet-50/ResNet-101 for 2D feature extraction, pre-trained on ImageNet [35, 36]. Research in [14, 31] utilised the pre-trained Deeplab v3+ [37] on the ADE20K dataset [38]. A recent study by Wang et al. [33] employed the Segformer [39], initialised with weights from ImageNet. It is noted that some studies have adopted iterative training with distinct learning rates for each input such as [30], while others opted for a singular global learning rate and consistent training settings such as optimisers and schedulers for parallel training across both input modalities [29, 31, 33, 40].

We observed that methods based on hybrid architectures with multiple inputs, such as RGB and geometry representations derived from TSDF, achieve better performance compared to models with single inputs in volume networks and view-volume networks due to multi features input. In this research, we extend our previous model DBNet in [9] by proposing MDBNet a hybrid model that simultaneously train on two distinct representations of the scene with F-TSDF and RGB inputs. We inspired by studies in [17, 19, 20], we employed hyperbolic tangent transformations on the identity features within our network and projecting RGB features from 2D to 3D using planner convolution layers.

However, we observe that SSC methods with perspective camera inputs are constrained by their limited input modalities and partial scene coverage, making them inadequate for applications requiring fully immersive VR environments. In the following section, we will discuss more about constructing 3D space with semantic completion including the occluded region from single full panorama view.

2.2 3D semantic scene completion (SSC) from single 360° view

Recent studies have extended 3D reconstruction to 360° inputs. For instance, ODGS [41] reconstructs 3D scenes represented by Gaussian splatting from multiple omnidirectional images, while AURORA [42] constructs indoor 3D spaces from sequences of RGB-D frames captured by a moving sensor. Only a few works address the full 3D reconstruction from a single 360° input. For example, Li et al. [43] employed CNNs for surface reconstruction from RGB-D, but did not generate annotated 3D models with semantic labels. Earlier methods such as DuLa-Net [44] mainly focuses on room layout estimation and struggle with occluded objects, while Meng et al. [45] proposed a method to recover the room structure without objects semantics from the RGB input. RepF-Net [46] detects objects in omnidirectional RGB images without performing 3D reconstruction and labelling. These approaches differ from our research objective in both the input/output format and the target representation.

On the other hand, the study by Kim et al. [47] employed SegNet [48] to extract scene semantics from 2D RGB inputs, generating a 3D model by mapping 2D points into 3D space using depth information. The resulting 3D point cloud is then grouped into clusters based on object labels, and block structures are reconstructed from these clusters using point occupancy to approximate the scene's geometry. In contrast, Kim et al. [49] proposed EdgeNet360, which is the most relevant to this research. It demonstrates densely annotated 3D models using depth-only 360° inputs. In that work, the authors infer 3D SSC from 360° depth input. However, we observe holes and incomplete objects in the reconstructed 3D SSC model, which reduce the scene fidelity in the VR space. Nevertheless, a gap remains in developing frameworks that integrate both RGB data and depth for fully annotated 3D SSC reconstructions with 360° coverage.

In this research, we extend the inference capabilities of the pre-trained MDBNet model that originally trained on densely annotated datasets of indoor perspective scenes. We adopt a spherical-to-cubic projection technique for RGB data and apply a 3D rotation method to depth. MDBNet is adapted to process 360° RGB-D inputs, enabling the generation of comprehensive 3D models suitable for immersive VR environments. While EdgeNet360 in [49] also produce detailed 3D reconstructions from depth-only inputs, our proposed framework bridges the existing gap in the literature by being adaptable to recent indoor SSC models pre-trained on both RGB and depth perspective views. This adaptability enhances its applicability to VR environments and facilitates further advancements in semantic scene completion.

2.3 Combining audio and visual data in 3D virtual space

Different methods have been introduced to model the properties of room acoustics, enabling the reproduction of spatial audio effects in virtual environments [49–51]. Several approaches existed for synthesising and generating RIRs [52], which can be broadly categorized into algorithmic methods, such as in [53–55], and deep learning methods, as in [56–60]. Some algorithmic methods, like [54] and [55], estimate RIRs in simplified or empty 3D scenes. In contrast, deep learning approaches increasingly leverage audio-visual inputs to estimate RIRs. However, both categories predominantly focus on RIR estimation without explicitly analysing the relationships between inferred 3D objects with semantic properties and the estimated RIRs. Consequently, there remains a gap in applying estimated RIRs to predicted 3D meshes for practical use.

Some studies investigated the theoretical relationships between 3D mesh surfaces and acoustic sound filed properties. For example, [61] demonstrated that surface features such as gaps and cracks significantly affect sound field reflections, causing localized increases in echo energy, with sensitivity affected by surface gap features such as smoothness, size, shape, and incident angle. Similarly, the study in [62] emphasized the critical role of edges in auralization, using edge diffraction models to simulate how sound bends around surfaces. The authours in that work identified four parameters wich are diffraction level, cutoff frequency, slope of the response, and phase of the diffraction to describe the sound behavior while still capture the main features of how sound reflects from small surfaces. Furthermore, the study in [63] showed that the perceptual impact of detailed diffusive surfaces such as triangular prisms on reverberance and spaciousness is noticeably stronger than flat surfaces. [49] further confirmed that voxelised 3D meshes result in better acoustic realism compared to simpler block-based models, building on earlier findings by [47]. Both studies [47, 49] used EDT and RT60 measurements to evaluate sound quality in VR environments for similar rooms.

Together, these findings highlight that the realism and perceptual accuracy of spatial sound also depend on the fine structural details of 3D surfaces. High-fidelity mesh reconstructions contribute to auditory realism and provide more consistent sensory experience across audio and visual cues, which are critical for immersive VR experiences.

Regarding sound rendering within VR environments, [47, 49] utilised Unity with sound spatialisation plug-ins: Google Resonance Audio [3] in [47] and Steam Audio in [49]. Notably, [49] found that Google Resonance Audio produced inferior audio quality when paired with ESS and voxel-based models. However, VR gaming engines have been widely adopted for creating immersive experiences in virtual spaces. Popular VR gaming engines include Unity, Unreal Engine [4], CryEngine [5], AppGameKit VR [6], ApertusVR [7], and Urho3D [8]. Among these, Unity and Unreal Engine are the most commonly used due to their community support, user-friendly interfaces, and

[3] https://resonance-audio.github.io/resonance-audio/(accessed in 2025)

[4] https://www.unrealengine.com/en-US(accessed in 2025)

[5] https://www.cryengine.com/(accessed in 2025)

[6] https://www.appgamekit.com/dlc/vr(accessed in 2025)

[7] https://apertusvr.org/(accessed in 2025)

[8] https://urho3d.io/(accessed in 2025)

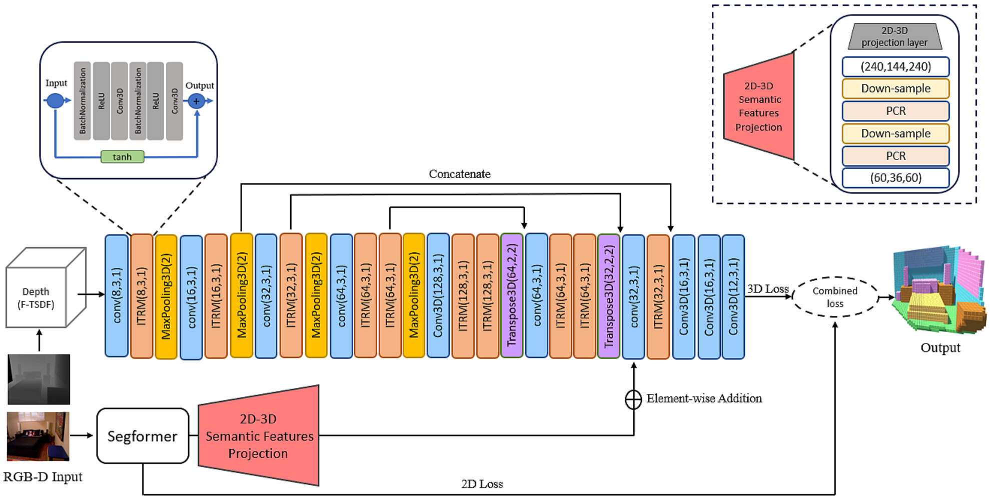


Fig. 2: MDBNet is a dual-head network that processes 2D RGB semantics via a pre-trained Segformer with 2D-3D projection and geometric data via a 3D CNN with ITRM blocks. The network optimises a combined loss, which is a weighted sum of 3D loss and 2D semantics loss.

advanced rendering capabilities [64, 65]. Unity is particularly noted for its lightweight build and ease of use compared to Unreal Engine [64, 66, 67]. Additionally, some studies have explored Unity for spatial sound experiences. For example, [68] proposed an optimised binaural sound rendering method for Unity using continuous-azimuth head related transfer functions (HRTFs) to improve localization accuracy. Similarly, [69] developed a VR orchestral concert experience using Unity combined with the Oculus spatialiser Native toolkit for audio specialization. The proposed framework in this research leverages Unity, and the Steam Audio Plug-in for advanced 3D sound spatialisation.

In this work, we analyse the quality of the rendered sound within the full 3D SSC by evaluating RIR acoustic parameters, such as EDT [70] and RT60 [71]. EDT is a metric used to evaluate the acoustics of adjacent reflectors by considering the energy carried by the early reflections [72, 73]. RT60 is related to the average absorption, location of room boundaries, and room size, describing reverberation from a physical perspective [72, 73].

3 Methodology

In this section, we present our method for predicting 3D SSC from perspective RGB-D input. We explain how we extend the inference capability of the pretrained MDBNet to handle a single full panoramic RGB-D input. Additionally, we describe our approach for room acoustic modelling to integrate spatial sound within the reconstructed 3D space to enhance immersion in VR environments.

3.1 3D SSC from a single RGB-D perspective input

We propose MDBNet deep learning neural network for 3D SSC prediction. The architecture of the proposed MDBNet is depicted in Figure 2. This model features a dual-head network, facilitating learning simultaneously from each network head within a single pipeline. The system processes each scene using two distinct modalities: a 2D input consisting of RGB image at a resolution of 640×480, and depth map data preprocessed as the form of F-TSDF for data representation within 3D space, which captures geometric information with dimensions of 240×144×240. We propose a modification to the residual blocks in our previous work [9]. The modification includes implementing identity transformation within full pre-activation residual module (ITRM) by adding a hyperbolic tangent (Tanh) function on the identity features within the residual blocks. The Tanh activation function is employed in various research contexts, particularly in scenarios where TSDF or SDF are used as input. Its primary purpose in such cases is to manage data distributions within a normalised range, aligning with the inherent data range of TSDF or SDF, as demonstrated in [74, 75]. In the F-TSDF representation, voxels in visible or empty spaces above surfaces are given values ranging from 1 to 0, while those in occluded areas have values from -1 to 0, creating steep gradients at objects surfaces [6]. The application of the Tanh function is particularly advantageous in this context, as it preserves the sign of the input with positive signals for visible space and negative ones for occluded regions, while normalizing the values to a range between [-1, 1]. In the domain of SSC, the Tanh activation function has been applied to part of identity features in a different context [17]. Our research extends this exploration by investigating additional context for the application of Tanh.

The model generates an output with a four-dimensional structure sized 60×36×60×12. The 12 channels represent the dataset classes ranging from 0 to 11. Class 0 is designated for empty spaces, whereas the remaining classes represent various object categories found in the NYUv2 [76] and NYUCAD [77] datasets, including ceiling, floor, wall, window, chair, bed, sofa, table, TV, furniture, and objects. Further details on this architecture will be discussed in the subsequent subsections.

3.1.1 2D semantic features

The incorporation of 2D RGB semantic features beside the F-TSDF features, can provide more guidance for SSC model learning. Specifically, RGB semantics add surface features to the objects in scenes, features that are absent in methods relying solely on depth maps as input. Transfer learning emerges as the most effective strategy for this adaptation process. It facilitates the efficient extraction of these RGB semantic features, enabling the system to benefit from learning more diverse features across larger dataset. Consequently, to optimise RGB input utilisation, we employ the Segformer 'B5' model, which is known for its superior accuracy and performance [39]. This Segformer model pre-trained on ImageNet and fine-tuned on the ADE20K dataset at a resolution of 640×640, leverages high-resolution image processing, aligning closely with the resolution of images in the NYU datasets [76, 77]. Given the limited size of the NYU dataset and its class overlap with ADE20K, it presents an ideal scenario for transfer learning. We adopted a transfer learning strategy by keeping the encoder's weights fixed and initialising the decoder's weights with those pre-trained on ADE20K, followed by fine-tuning on the NYU datasets [33].

Features extracted from 2D RGB images are projected and mapped onto the corresponding coordinates in 3D space by taking advantage of the existing depth map input. Aligned with the projection method described in [20], we utilised the depth values from the depth image, along with the intrinsic camera matrix and the extrinsic camera matrix to project a pixel from the 2D image plane to a 3D point, we map 2D features into scene surfaces in the 3D space. Then, these volumetric surface features are fused with the F-TSDF input within 3D network branch.

Different fusion methods based on element-wise addition are implemented to assess the model's performance, including early, middle, and late fusions [13]. The aim of investigating different fusion methods is to identify the best location to add the projected RGB semantic features into the geometric information represented by F-TSDF within the network.

In early fusion, the full-resolution projected 3D surface features $(240 \times 144 \times 240)$ are combined with the F-TSDF input before entering the 3D network branch. This enables the network to jointly learn from both modalities from the beginning with the unrefined nature of the early features. For middle and late fusions, the projected 3D surface features downsampled to align with the resolutions of the network's intermediate $(15 \times 9 \times 15)$ and later $(60 \times 36 \times 60)$ layers, respectively. This downsampling process employed the Planar Convolution Residual (PCR) block [19], a variant of the Dimensional Decomposition Residual (DDR) block [22], which breaks down the standard 3D convolution into three sequential one-dimensional layers along three orthogonal axes. The PCR uses planar convolutions with kernel dimensions where one of the three sizes is 1, preserving the planar characteristics of the 3D scene and reducing the parameter count relative to standard residual blocks. In middle fusion, the network fuses the RGB semantic features in the latent space within the 3D network branch, where the representations are coarse spatially. In contrast, late fusion combines the RGB semantic features at a more refined stage of F-TSDF features within the 3D network branch compared to early and middle strategies, allowing RGB semantic features to guide the final 3D semantics reconstruction. The late fusion method produced the best performance among the three strategies in our experiments.

3.1.2 Combined loss function

We supervise the two inputs of MDBNet jointly using a combined loss function that merges the 2D semantic loss and the 3D loss for SSC, employing a weighted sum approach. This method utilises a weighting parameter λ to balance the contributions of the two losses, designated as L_{SS} for 2D semantic loss and L_{SSC} for the 3D SSC loss defined in our previous work DBNet [9]. The combined loss function is formulated in the following Equation 1:

$$L = \lambda L_{SS} + L_{SSC}. \tag{1}$$

Aligned with [33], we employ the smooth cross-entropy loss, denoted as L_{SS}, to measure the loss for 2D RGB semantic predictions. The L_{SSC} trains the model with F-TSDF 3D features after integrating the projected 2D RGB semantic features in the current context. It employs a smoothed weights over the different classes in the dataset through an unsupervised clustering algorithm, K-means. That re-weighting approach address the imbalance between occupied and empty voxels and effectively handles imbalances across different classes within the occupied voxels in the dataset. In our previous work, DBNet [9] L_{SSC} combines the benefits of re-sampling and class-sensitive learning to address the inherent class imbalance in the data. It employs a smoothed weights through an unsupervised clustering algorithm, K-means. The computation of L_{SSC} loss assesses the discrepancy between the predicted label p and the genuine label y across the voxels of a scene A. For each voxel v within A, the predicted and actual labels for a given voxel v are indicated by p_v and y_v, respectively. Each voxel label is assigned a specific weight w_v using the reweighing method based on K-means clustering. The loss function is defined as follows in Equation 2:

$$L_{SSC}(p, y) = -\sum_{v=1}^{A} w_v \,.\, y_v \,.\, logp_v. \tag{2}$$

3.2 Extend MDBNet to MDBNet360

We extend MDBNet's inference capabilities to 360° RGB-D data by incorporating spherical-to-cubic projection and 3D transformation for comprehensive 3D reconstruction with 360° surroundings. The proposed design generates cubic views from 360° RGB-D input by converting the spherical RGB data into six perspective images. Following [78].

To compute the F-TSDF from the spherical depth map, depth grids are first generated for each cubic view. Point clouds are derived from the spherical depth data. We establish a mapping between 3D depth points and their corresponding pixels in equirectangular images. This mapping follows the general principles of spherical-to-Cartesian transformation, as implemented in prior works such as [49]. The Cartesian coordinates (x, y, z) are calculated using the latitude and longitude from the equirectangular image.

An occupancy voxel grid is constructed to represent the scene's surface. This is achieved by simulating four perspective views (left, front, right, and back) with each view rotated 90° around the Y-axis. Because spherical projection introduces polar distortions and uneven sampling near the poles (top and bottom) [49, 79, 80], and because training data rarely include ceiling (+90°) or floor-only (−90°) viewpoints, these regions are replaced with fitted planes in our final 3D design and excluded from F-TSDF processing and prediction by MDBNet. The transformation of Cartesian coordinates (x, y, z) for each view is performed using the following rotation matrix:

$$\begin{bmatrix} x' \\ y' \\ z' \end{bmatrix} = \begin{bmatrix} \cos(\theta) & 0 & \sin(\theta) \\ 0 & 1 & 0 \\ -\sin(\theta) & 0 & \cos(\theta) \end{bmatrix} \cdot \begin{bmatrix} x \\ y \\ z \end{bmatrix}. \tag{3}$$

The F-TSDF is then calculated for each 3D view. The TSDF value represents the Euclidean distance of each voxel to the nearest surface voxel using specific truncation threshold t to reduce both computational load and memory usage within the perspective cubic view. The TSDF is flipped to provide strong gradients on surface [6] :

$$F\text{-TSDF} = \text{sign}(TSDF) \cdot (TSDF_{\max} - |TSDF|). \tag{4}$$

The sign in Equation 4 provides information about whether the voxel is in front of or behind the object's surface. In the F-TSDF representation, voxels in visible or empty spaces above surfaces are assigned values ranging from 0 to 1, while those in occluded areas are assigned values from -1 to 0, resulting in steep gradients at object surfaces. Then we pass the RGB perspective view with corresponding F-TSDF inputs into the proposed model. We construct a comprehensive inference pipeline by combining predictions from multiple MDBNet inferences. Our proposed architecture generates four 3D volumes, with boundary overlaps occurring between adjacent 3D views. These views are merged within a single comprehensive view using the summation rule [81] as illustrated in Figure 3. The MDBNet's outputs in the overlapping regions are aggregated using summation. For each voxel with output P_{ij} for class i predicted by MDBNet classifier j, the total sum of the values for class i across all m classifiers is calculated as follows:

$$O_i = \sum_{j=1}^{m} P_{ij}. \tag{5}$$

$$C = \arg\max_i (O_i). \tag{6}$$

Post-processing is applied to all inferred 3D views, including fitting planes (walls, ceiling, and floor) in the room to enhance overall scene quality, ensuring a more coherent and visually realistic representation.

The 3D room, with the aggregated views, is then imported into Unity with Steam Audio for semantic-level material assignment and sound rendering. Steam Audio provides 11 acoustically predefined materials, including wood, carpet, glass, plaster, metal, and ceramic. These materials are assigned to objects in the scene following the method described in [49]. Consequently, the sound is rendered using the materials corresponding to the scene semantics, thereby providing a spatially realistic audio experience. The acoustic modelling process is described in more details in the following section.

3.3 Room acoustic modelling

In this research, we use the Steam Audio plug-in with Unity to render sounds within the 3D volumes generated by MDBNet360 in virtual space.The RIR of the virtual space is measured by playing an ESS signal from a single virtual sound source and recording the response at the listener position. To generate ESS audio, we follow the approach proposed by Farina [82–84], utilising Equation 7:

$$x(t) = \sin\left[\frac{\omega_1 \cdot T}{\ln\left(\frac{\omega_2}{\omega_1}\right)} \cdot \left(e^{\frac{t}{T}\cdot\ln\left(\frac{\omega_2}{\omega_1}\right)} - 1\right)\right]. \tag{7}$$

The virtual sound source sweeps through the samples t of the exponential sine signal x(t), starting from the lowest angular frequency ω_1 and progressing to the

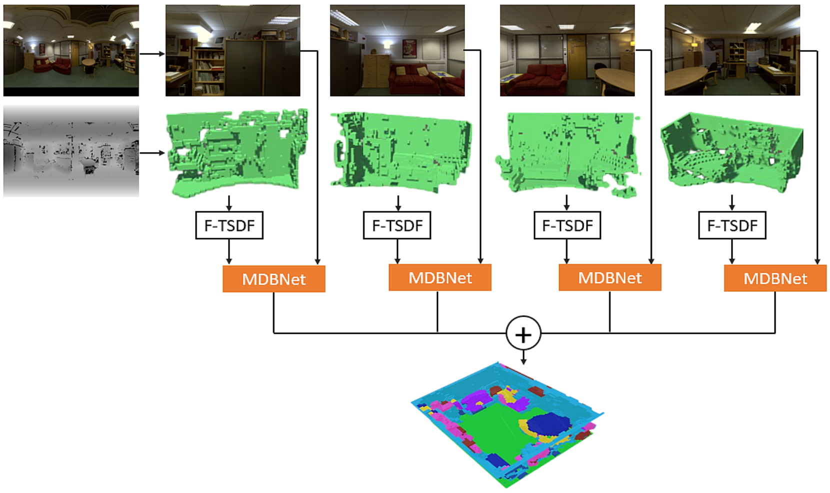


Fig. 3: MDBNet360: RGB-D projection and prediction on full panorama MR scene from CVSSP dataset using MDBNet SSC model.

highest angular frequency ω_2, as depicted in Equations 8 and 9, respectively. The sweep has a duration of T.

$$\omega_1 = 2 \cdot \pi \cdot f_1/fs \tag{8}$$

$$\omega_2 = 2 \cdot \pi \cdot f_2/fs \tag{9}$$

The RIR is extracted from the recorded sound at the listener and saved in WAV format. Next, we measure the room acoustics parameters, including RT60 and EDT. To estimate RT60, we analyse the room's RIR and calculate the time it takes for the sound to decay by 60 dB, as defined by ISO 3382-1:2009 [85]. This approach employs a linear least-squares fit to determine the slope between 0 dB and -60 dB [71, 86]. EDT is estimated using the slope of the decay curve, determined from the fit between 0 and -10 dB. The decay time is then calculated from the slope as the time required for a 60 dB decay [70, 86]. The values are averaged for both EDT and RT60 across six octave bands, ranging from 250 Hz to 8000 Hz, to ensure comparability with previous methods using similar bands [47, 49]. In order to assess the perceptual relevance of the observed discrepancies in EDT and RT60 values, we define their just noticeable differences (JNDs). According to recommendations from the literature, the JND thresholds are set at 20% for RT60 [87] and 5% for EDT [88].

4 Experiments

4.1 3D SSC using MDBNet

In this section, we introduce the implementation and experimental settings of MDBNet model to generate 3D SSC scenes from perspective RGB-D input.

4.1.1 Training and validation

We conduct our experiments using the PyTorch framework, on a single Nvidia RTX 8000 GPU. Both 2D and 3D network branches are trained simultaneously with MDBNet. Due to the two types of input representation, we employ different learning rates to achieve effective performance as demonstrated in [89]. For the 2D input modality (RGB), we employ a pre-trained Segformer model [90], which is fine-tuned on the ADE20K dataset [38] at an image resolution of 640 × 640. In the pre-trained model, we keep the encoder's weights fixed and fine-tuned the decoder layers, starting with a learning rate of 1×10^{-4}. Following the study by Wang et al. [33], we used the AdamW optimiser with 0.05 weight decay, and learning rate governed by a cosine decay policy, starting from the initial value and decreasing to a minimum of 1×10^{-7}. For the 3D input modality, we opt Stochastic Gradient Descent (SGD) with a momentum of 0.9 and a weight decay of 5×10^{-4}. The OneCycleLR scheduler is utilised to adjust the learning rate, beginning at 0.01 [9]. We train the MDBNet model for 100 epochs, with batch sizes set to 4 for training and 2 for validation. To mitigate the risk of overfitting on the training dataset, we incorporate an early stopping as a regularization method [91] with a patience setting of 15 epochs. In our loss function, we experiment with a coefficient λ set to 1 and normalised the scale of L_{SS} to match that of L_{SSC} by setting λ to 0.5. The model exhibits stability across both configurations and demonstrates effective learning. Although the score ranges for both settings show considerable overlap, a slightly higher SSC score is observed with $\lambda = 1$, achieving 60.1 ± 1.0 compared to 59.2 ± 1.3 with $\lambda = 0.5$. Furthermore, to ensure the performance reliability of our results, we implement K-fold cross-validation [92–94], dividing the training set into three folds at random, and preserving the weights from each fold for subsequent evaluation on the test set, thereby quantifying the model's performance uncertainty.

4.1.2 Datasets

Our research leverages the NYUv2 and NYUCAD datasets as benchmarks for conducting our experiments. NYUv2 consists of 1449 realistic RGB-D indoor scenes captured via a Kinect sensor with a resolution of 640×480. The datasets are divided into 795 training instances and 654 testing instances. However, as discussed in [6], there is some misalignment between the depth images and the corresponding 3D labels in the NYUv2 dataset, which makes it difficult to evaluate accurately. To address this problem, we also use the high-quality NYUCAD synthetic dataset, which projects depth maps from ground truth annotations and avoids misalignment.

4.1.3 Metrics

We adopt Precision, Recall, and IoU as the evaluation measures for the SSC, following the approach of Song et al. [6]. For the semantic scene completion task, both the observed surface and occluded regions are evaluated. We present the mIoU scores for semantic classes, excluding the empty class. In the scene completion task, all non-empty voxels are classified as '1', while empty voxels are labeled as '0'. The binary IoU is computed for the occluded regions in the view frustum along with precision and recall measures. We have observed that there is no standardized method for selecting

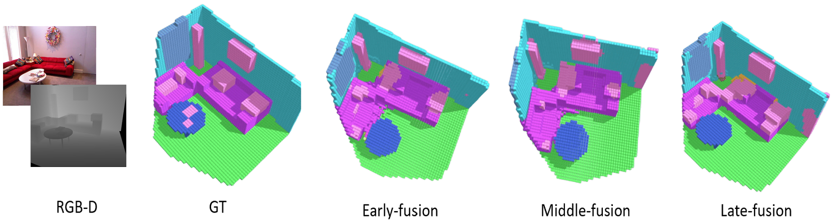


Fig. 4: Illustration of model output using different fusion methods on NYUCAD dataset

the scene completion area, leading to slight variations among researchers in the field. Some researchers, as seen in [20] select the occupied occluded voxels while the empty occluded voxels are re-sampled. On the other hand, SPAwN [32] bypasses re-sampling step for empty occluded voxels and evaluates all unoccupied voxels. Other studies, such as PALNet [28], DDRNet [22], and AICNet [23], include all occupied voxels in the scene, combining visible surfaces with occluded regions for scene completion evaluation. In this research, we follow [20] by evaluating all occluded occupied voxels and re-sampling empty occluded ones. As highlighted in [23, 95], the mIoU metric is considered more critical than IoU. Nonetheless, the results for all metrics are average across K-fold cross-validation to derive the final scores.

4.1.4 Ablation study

In this section, we conduct ablation studies on the NYUCAD dataset to evaluate the effectiveness of our proposed RGB feature fusion methods and the various components of our model design.

The model with the proposed combined loss function is trained using various methods to fuse the 3D projected RGB semantic features. The results, as reflected within average scores presented in Table 1, indicate that our model is capable of learning effectively using these different fusion strategies as shown in Figure 4. Among them, the late fusion method demonstrates the best average score, with the highest stability in performance, as evidenced by the lower standard deviation scores that indicate less uncertainty in performance. Specifically, we observe that the TV object is not well recognised in some folds when using the early and middle fusion methods, whereas it is consistently recognised across all folds with the late fusion approach. Consequently, we select the late fusion approach for RGB semantic features to further evaluate the model's performance across different components. On the other side, to confirm the impact of each component within our MDBNet, we modify the previous DBNet model [9] by integrating new components and conduct comprehensive experiments to evaluate their contributions, as detailed in Table 2. Initially, we train our model with RGB-D input and apply our combined loss, which includes our proposed re-weighting 3D loss [9], achieving SSC score of 59.0%. In the second experiment, we replace our proposed re-weighted loss [9] with a re-sampling-based loss from [6]. This substitution results in a significant decrease of 6.5 percentage points (pp) in the SSC score, underlining the critical role of both RGB features and our re-weighted 3D loss within

Table 1: Ablation studies using different RGB features fusion methods.

Fusion Method	SC-IoU%	SSC-mIoU%
Early	**80.5**±1.0	57.1±2.3
Middle	79.3±0.9	55.8±2.5
Late	79.3±0.6	**59.0**±0.1

Table 2: Ablation studies on the NYUCAD dataset evaluating MDBNet components with RGB-D input.

Method	SC-IoU%	SSC-mIoU%
$L_{ss} + L_{SSC}$ (re-weighting)	79.3±0.6	59.0±0.1
$L_{ss} + L_{SSC}$ (re-sampling)	**80.5**±0.9	52.5±0.9
$L_{ss} + L_{SSC}$ (re-weighting) + ITRM	79.8±0.8	**60.1**±1.0

the proposed combined loss function and their impact on the model's performance. In the third experiment, we apply our combined loss function and enhance the 3D branch of MDBNet by replacing the original residual blocks in DBNet with the proposed ITRM blocks. This modification leads to further improvements, achieving an SSC score of 60.1%, which is a 7.6 pp increase over the second experiment's score of 52.5%. Although the score is slightly higher than that of the first experiment (59.0%), suggesting a positive trend, the difference lies within the standard deviation is not statistically significant. Figure 6 provides a visualization on different components outputs.

4.2 3D audio-visual VR scenes production using MDBNet360

We generate 3D scenes with 360° surroundings using the MDBNet360 model as described in Section 3.2. MDBNet360 preprocesses RGB spherical images to produce cubic perspective views and combines them with F-TSDF 3D data, using a truncation value set to 0.24 meters. To infer the 3D volumes, we utilise the saved weights of the pre-trained MDBNet model on the NYUCAD dataset. The average inference time to produce a full 3D room is 2.57 minutes on a single NVIDIA RTX 8000 GPU. The model utilises a 0.02 meter voxel size within a grid of 240 × 144 × 240 for scene input representation, which is scaled down to 60 × 36 × 60 for output, to remain compatible with the MDBNet settings. This representation covers room sizes within 4.8 × 2.88 × 4.8 meters. We test the proposed method using the CVSSP dataset [9]. The CVSSP dataset consists of five indoor scenes with 360° RGB-D and ground-truth acoustic parameter measurements [47, 49]. For our simulations, three scenes are selected: the Meeting Room (MR), Kitchen (KT), and Usability Lab (UL). The Listening Room (LR) and Studio Hall (ST) are excluded. The LR is omitted because it contains acoustically controlled materials, which would not provide relevant results

[9] http://3dkim.com/research/VR/index.html(accessed in 2025)

Table 3: Material assignment table for objects.

Object	Steam Audio Material
Ceiling	Wood
Floor	Carpet
Wall	Plaster
Window	Glass
Bed	Carpet
Sofa	Carpet
Chair	Wood
Table	Wood
TV	Glass
Furniture	Wood
Object	Metal

for our study. The ST is excluded due to its dimensions being significantly larger than those used for constructing the 3D voxels. We enhance the depth data following the method described in [49].

After generating the 3D rooms using MDBNet360, we import the 3D models into Unity, which is integrated with the Steam Audio plug-in for 3D audio-visual rendering, enabling an immersive VR environment.

4.2.1 Room acoustics estimation

In each scene within the Unity platform, a virtual sound source and listener are positioned to align with the ground-truth locations. Unified simulation settings are applied across all scenes. The semantic information obtained from MDBNet360 is utilized in the VR system to assign material properties and acoustic parameters to objects in the reconstructed 3D scene. For instance, a corresponding Steam Audio Geometry material is mapped to each object category. Table 3 lists the objects and their corresponding materials as described in [49], this mapping serves as a rough estimation intended to generate plausible sounds and reproducible simulations that allow fair comparison with existing SOTA methods [47, 49]. Before rendering the sound, the scene must be saved and exported to ensure that all effects, including the geometry materials applied to each component, are correctly integrated.

Following the ground truth, where both the sound source and listener are static, we design the simulations using static settings with precomputed, or ‘baked’ effects to reduce CPU usage. An empty game object is added to each scene to assign the Steam Audio Probe Batch, which creates sound probes. These probes serve as points where Steam Audio measures reflections and reverberation during the baking process. At runtime, the relative positions of the source and listener to the probes are used to quickly estimate these acoustic effects. Additionally, for the virtual sound source in the scene, we attach the ESS audio file generated based on the method described in Section 3.3. The ESS audio is generated with a sampling rate of 48,000 Hz and saved at 16-bit. The ESS audio with frequencies ranging from 20 Hz to 20,000 Hz, is rendered with Steam Audio geometry materials within each virtual room. To generate spatialise sound, we choose the spatialise option and set the Spatial Blend to the 3D to generate immersive rendered sound. For the Steam Audio Source we apply HRTF-based binaural rendering, utilising the default Nearest interpolation option to control how HRTFs are adjusted as the sound source moves relative to the listener. The impact of HRTF is more pronounced in scenarios that involve moving sound sources or listeners. Distance Attenuation is applied to the Steam Audio Source, considering the Spatial Blend setting. If the Spatial Blend is set to 2D, Distance Attenuation is effectively disabled. A Physics Based distance attenuation model is employed, where the volume curve and other curves defined in the 3D sound settings of the Audio Source are disregarded. This differs from the curve-driven attenuation model, which is controlled by the volume curve specified in the Audio Source settings. We choose the Attenuation Settings to be with Air Absorption to apply frequency-dependent calculations for air absorption effects. The Simulation Defined option is chosen, which specifies how the air absorption values are determined using exponential decay pattern, where higher frequencies diminish more rapidly over distance compared to lower frequencies. Furthermore, reflections from the surfaces that reach the listener are simulated by choosing the Reflection option. These reflections are processed with HRTF and baked at the static listener. At this stage, the scene is saved and exported. Additionally, we attach the Steam Audio Baked Listener and Steam Audio Listener, with simulated reverberation, to the Audio Listener in the virtual room. The influence radius is adjusted based on the room size. The sound is baked at the Audio Listener, and after that, the final effects are saved and exported.

To measure the RIR, we play the ESS sound at the source position in the scene and record the rendered ESS sound at the listener. The recorded sound is then convolved with the ESS inverse filter to extract the RIR. Then, we measure the average EDT and RT60 acoustic parameters among the six octave bands as described in Section 3.3.

5 Results

5.1 3D SSC from a single RGB-D perspective input using MDBNet

To evaluate the performance of MDBNet using the NYUv2 and NYUCAD datasets. Quantitative comparisons of MDBNet results with SOTA approaches are detailed in Tables 4 and 5. Unlike previous studies, which did not specify the performance uncertainty, we averaged our scores across three folds to more accurately represent generalisation performance and to ensure an unbiased assessment. Due to the variations in how researchers select the scene completion area, as discussed in Section 4.1.3, these differences do not necessarily show true performance gaps between SOTA models. Also, [23, 95] highlight the importance of mIoU over IoU. However, for a fair comparison, we focus on semantic scene completion, which relates to the object area and is measured using standardized criteria.

We compare MDBNet with SOTA methods that utilise hybrid architectures, focusing on voxel-based semantic segmentation on the NYUv2 dataset, as shown in Table 4. Our approach significantly outperforms current SOTA models, achieving a remarkable increase in mIoU scores by 3.1 pp and 2.7 pp over the previously leading methods,

Table 4: Results on the NYUv2 dataset include averages and standard deviations for Precision, Recall, IoU, and mIoU metrics. In the input column, 'D' means depth map only. In the method column, '*' represents the view-volume architecture type.

Method	Input	Res.	Scene Completion (SC)			Semantic Scene Completion (SSC)											
			Prec.	Recall	IoU	ceil.	floor	wall	win.	chair	bed	sofa	table	tvs	furn.	objs	mIoU
AMMNet$_{Segformer}$[96]	RGB-D	(60,60)	90.5	82.1	75.6	46.7	94.2	43.9	30.6	39.1	60.3	54.8	35.7	44.4	48.2	35.3	48.5
CleanerS[33]	RGB-D	(60,60)	88.0	83.5	75.0	46.3	93.9	43.2	33.7	38.5	62.2	54.8	33.7	39.2	45.7	33.8	47.7
SISNet(voxel)[30]	RGB-D	(60,60)	87.6	78.9	71.0	46.9	93.3	41.3	26.7	30.8	58.4	49.5	27.2	22.1	42.2	28.7	42.5
PCANet*[19]	RGB-D	(240,60)	89.5	87.5	78.9	44.3	94.5	50.1	30.7	41.8	68.5	56.4	32.6	29.9	53.6	35.4	48.9
SPAwN[32]	RGB-D	(240,60)	82.3	77.2	66.2	41.5	94.3	38.2	30.3	41.0	70.6	57.7	29.7	40.9	49.2	34.6	48.0
DBNet (Ours)[9]	D	(240,60)	79.3±1.0	83.3 ±0.8	68.1±0.5	48.9	92.8	49.2	0.0	31.7	61.4	56.1	29.2	0.0	33.9	19.3	38.4±0.2
MDBNet (Ours)	RGB-D	(240,60)	80.3±3.7	81.8 ±6.5	67.6±2.1	47.2	92.6	49.9	47.6	46.8	66.2	62.1	37.1	35.7	45.2	36.9	51.6±1.5

Table 5: Results on the NYUCAD dataset include averages and standard deviations for Precision, Recall, IoU, and mIoU metrics. In the input column, ‘D’ means depth map only. In the method column, ‘*’ represents the view-volume architecture type.

Method	Input	Res.	Scene Completion (SC)			Semantic Scene Completion (SSC)											
			Prec.	Recall	IoU	ceil.	floor	wall	win.	chair	bed	sofa	table	tvs	furn.	objs	mIoU
AMMNet$_{Segformer}$[96]	RGB-D	(60,60)	92.4	88.4	82.4	61.3	94.7	65.0	38.9	58.1	76.3	73.2	47.3	46.6	62.0	42.6	60.5
SISNet(voxel)[30]	RGB-D	(60,60)	92.3	89.0	82.8	61.5	94.2	62.7	38.0	48.1	69.5	59.3	40.1	25.8	54.6	35.3	53.6
SPAwN[32]	RGB-D	(240,60)	84.5	87.8	75.6	65.3	94.7	61.9	36.9	69.6	82.2	72.8	49.1	43.6	63.4	44.4	62.2
PCANet*[19]	RGB-D	(240,60)	92.1	84.3	86.3	54.8	93.1	62.8	44.3	52.3	75.6	70.2	46.9	44.8	65.3	45.8	59.6
DBNet (ours)[9]	D	(240,60)	86.5±0.9	91.1±1.1	79.6±0.1	66.7	93.6	60.7	15.7	51.4	68.9	68.7	45.6	0.0	44.9	29.3	49.6±1.2
MDBNet (Ours)	RGB-D	(240,60)	85.0±1.7	93.0±1.2	79.8±0.8	67.4	93.6	64.1	52.4	59.5	72.5	69.3	45.0	41.5	53.1	42.4	60.1±1.0

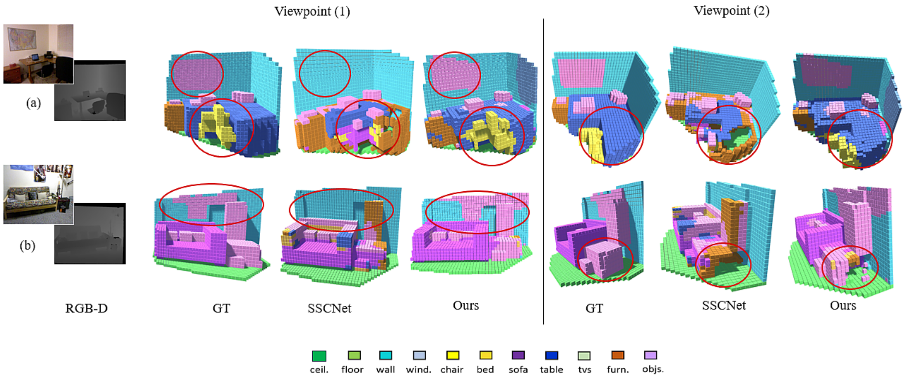


Fig. 5: Comparison of SSC results on the NYUv2 dataset: SSCNet (depth maps) vs. MDBNet (RGB-D). Objects are colour-coded, with circles marking key differences between GT and predictions.

AMMNet$_{Segformer}$ [96] which employed Segformer pretrained model for 2D RGB features, and PCANet [19], respectively. The efficacy of MDBNet is validated on the NYUCAD dataset, as shown in Table 5. Our average performance is competitive with AMMNet$_{Segformer}$ [96], and surpasses it when considering the upper-bound results. Furthermore, although our design surpasses SPAwN [32] on the NYUv2 dataset, it demonstrates performance comparable to the more resource-intensive SPAwN model, which utilises semantics priors calculated using surface normals.

On the other hand, we provide qualitative analysis to illustrate the effectiveness of MDBNet's components. To highlight the performance of MDBNet design and its success in generating more precise predictions, we present a series of visual comparisons using the NYUv2 dataset, as illustrated in Figure 5. These comparisons, made between our method and SSCNet [6], demonstrate the improved prediction accuracy offered by our approach. We achieve enhanced scene completion, particularly in the occluded parts of the scenes, as demonstrated in (a) and (b) of Figure 5. Additionally, by extracting semantic features from the RGB inputs, MDBNet exhibits superior performance, even surpassing the ground truth (GT) 3D volumes in certain regions. For instance, in Figure 5 (a), the RGB image shows both object and window existing on the walls. Our model successfully predicts the object and window voxels on the walls where they are absent in the GT 3D volumes. Figure 6 presents various scenarios within the NYUCAD dataset, comparing when our combined loss function uses weighting based on re-sampling [6] within the 3D loss, when it applies our proposed re-weighting [9], and when employing re-weighting [9] and incorporating ITRM representing the MDBNet model. The incorporation of class re-weighting in our combined loss significantly enhances the model's ability to identify underrepresented classes, such as TVs and chairs, as shown in Figure 6 in (a), (c), and (d). Additionally, our MDBNet offers better recognition of chairs with various shapes in the same figure in (b), (c), and (d), and it ensures enhanced differentiation between tables and chairs, as evident in (b) and (c). MDBNet model effectively recognises challenging classes like

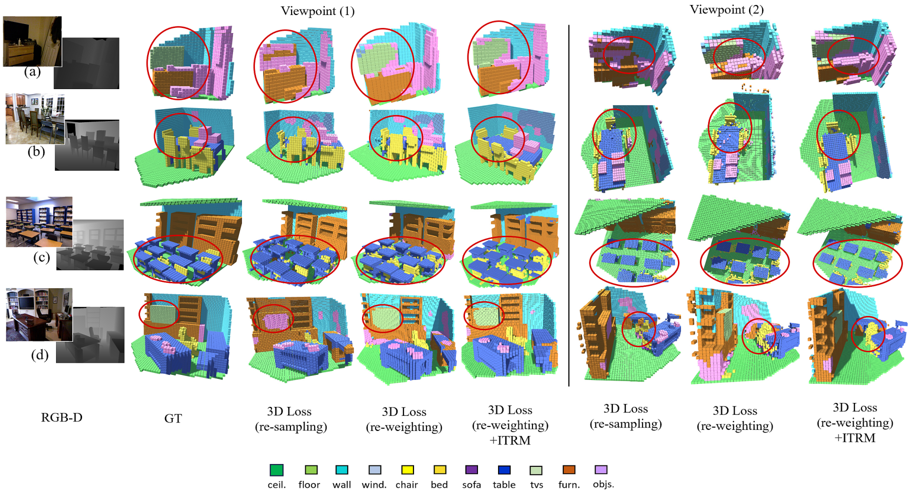


Fig. 6: SSC results with different components on NYUCAD dataset. From left to right: (1) RGB-D input; (2) GT; (3) combined loss with re-sampling; (4) combined loss with re-weighting; (5) combined loss (using re-weighting) with ITRM blocks. Objects are colour-coded, with circles highlighting key differences between GT and predictions.

windows and TVs, showcasing its robustness and adaptability. Additional results are available on our GitHub account: https://github.com/MonaIA1/Repo.

5.2 3D SSC from a single 360° RGB-D input using MDBNet360

The original MDBNet demonstrated superior results, significantly outperforming other SSC models. Due to the lack of ground truth 3D annotated data within CVSSP, we qualitatively assess the 3D voxelised models of the reconstructed rooms generated by MDBNet360. These models are compared with those produced by EdgeNet360 [49] an extension of EdgeNet [18]. We can clearly observe that MDBNet360 outperforms EdgeNet360 in semantic scene completion across all selected scenes from the CVSSP dataset. Notably, even with the low resolution of depth maps in the CVSSP dataset, where depth values are stored with 8-bit, which leads to a loss of fine object details, MDBNet360 exhibits a clear improvement in predicting and completing key scene components. For evaluation, we focus on objects that play a central role in understanding room structure and functionality, namely sofas, chairs, and tables. These elements were chosen because they are among the most commonly used indoor objects and influence spatial perception. To provide our qualitative comparison, we select a viewpoint that prominently displays these key objects, ensuring a clear visualisation of the model's reconstruction capabilities. As illustrated in Figure 7, MDBNet360 offers more detailed and complete representations of tables and chairs in the MR and KT scenes, where EdgeNet360 often struggles. For example, EdgeNet360 produces a partially reconstructed table in the MR scene, missing chairs in the room, and the

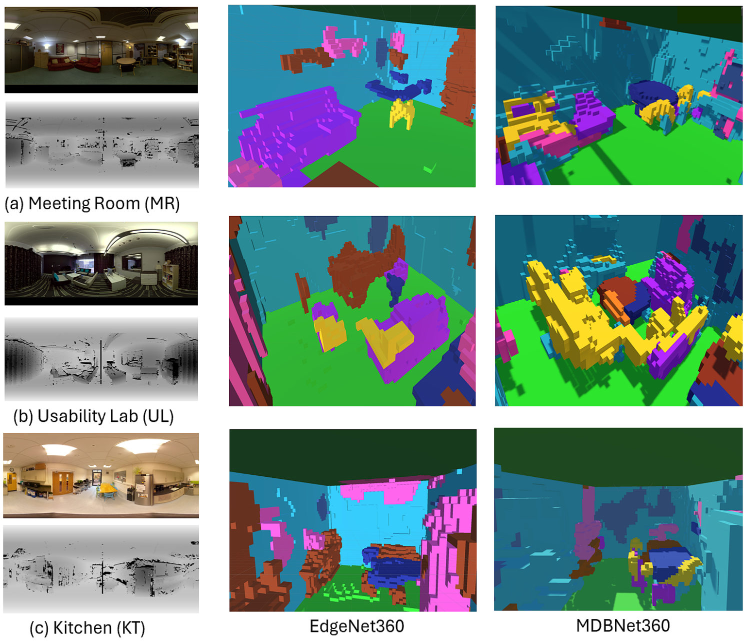


Fig. 7: Qualitative comparison between MDBNet360 and EdgeNet360 on three scenes in CVSSP data. From top to bottom: MR, UL, and KT.

omission of chairs around the table in the KT scene. Such inconsistencies negatively impact the spatial understanding of the room. In contrast, MDBNet360 maintains the structural integrity of the scene, improving geometric consistency. In the UL scene, EdgeNet360 fails to reconstruct the central table, significantly altering the perception of the room's layout. In addition, large portions of the sofas are missing, reducing the completeness of the scene. MDBNet360, however, preserves these crucial spatial elements, enhancing both the functional interpretation and the visual coherence of the scene. Furthermore, one of the key strengths of MDBNet360 is its ability to predict challenging scene features, such as windows and glossy doors, which are often difficult to detect and reconstruct due to their reflective properties and transparency. Despite some boundary errors, MDBNet360 successfully predicts the correct locations of these objects in both the UL and KT scenes. In contrast, EdgeNet360 exhibits significant semantic errors in estimating these objects, often either completely missing or misplacing them in its reconstructions. This performance disparity is largely due to MDBNet360's incorporation of features from dual inputs: RGB and depth, compared to EdgeNet360's reliance on solely on depth data. Nonetheless, our results indicate that MDBNet360 improves the completeness and fidelity of indoor scene reconstruction, particularly in the representation of essential structural elements which is an aspect crucial for high-quality semantic scene completion.

5.3 Spatial audio within virtual space

To provide a comprehensive evaluation of the virtual space, we assess the sound quality within the virtual rooms generated by MDBNet360. Specifically, we evaluate the RIR based on the EDT and RT60 acoustic parameters. Our results are compared with the ground truth (black line) measurements obtained from sound modeled in real space, and SOTA models Kim19 [47] and Kim20 [49]. We also provide the JND thresholds (dotted line) within our results to assess whether the observed differences in the EDT and RT60 values are likely to be perceptually noticeable. The thresholds are based on the ground truth set at 20% for RT60 [87] and 5% for EDT [88], as described in Section 3.3. In general, our approach demonstrates better performance in both EDT and RT60 compared to Kim19 and Kim20, as shown in Figure 8. In the figure, the EDT scores for our model in the MR and UL scenes outperform those of Kim19 and Kim20, being closer to the ground truth. However, for the KT scene, the EDT score predicted by MDBNet360 is slightly shorter than the ground truth. We attribute this discrepancy to errors in the 3D semantics, where cabinets are mislabeled as the wall voxels with plaster material. This mislabeling likely occurred due to inaccuracies in depth perception and the similarity between the cabinet colour and the wall colour in the RGB image, making it challenging for our model to accurately distinguish the cabinets. In the real world, cabinets typically have lower absorption coefficients than plaster walls, as their materials are more reflective. In the 3D voxel scene within Unity, the materials do not perfectly match the acoustic properties of their real world counterparts. Since the cabinets are labeled as wall voxels, they are assigned to plaster-like material properties. We observe some artifacts that affected the acoustic modelling, resulting in excessively high RT60 values exceeding thirteen seconds in UL scene only. These artifacts are likely caused by the presence of objects between the sound source and the listener (a situation not present in the MR and KT scenes) which are inaccurately modeled and assigned incorrect material properties. The high sensitivity of the sound listener likely contributes to this issue, as it could detect even minor sound reflections and scattering from the voxel model surfaces such in [4]. This can be considered as a technical limitation of Steam Audio, the spatial audio rendering plug-in. This can be avoided by slight adjustment of the listener's position and fine-tuning of simulation parameters, such as the Reflection Mix Level, which helps to reduce the artifacts and provides more reliable results.

However, the final VR space reconstructed by MDBNet360 demonstrates improved performance in both 3D visual scene prediction and spatial sound rendering compared to existing approaches. The rendered sound results are shared via Github account at: https://github.com/MonaIA1/Repo360.

5.4 Discussion

In this work, we address SSC problem, which involves the simultaneous determination of volumetric occupancy and object classification from a single RGB-D input, offering a limited perspective. We propose MDBNet, which provides an effective solution through the implementation of several components, including our combined loss function, the investigation of RGB fusion placement, ITRM blocks, and benchmark

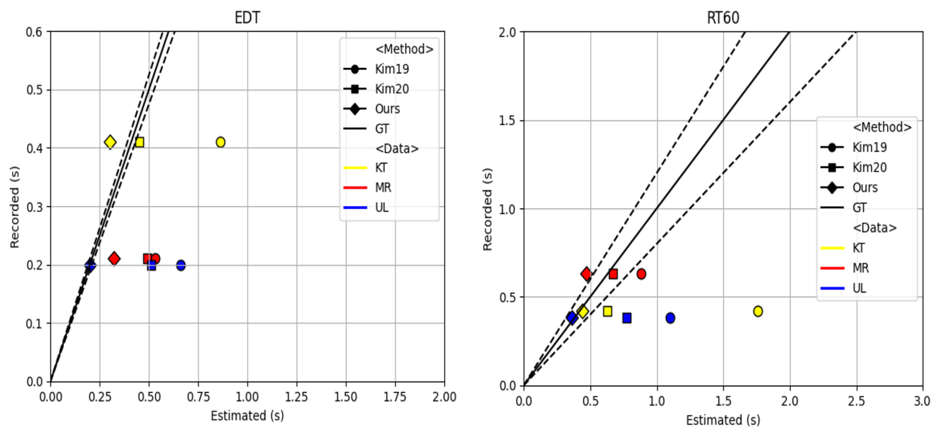


Fig. 8: EDTs and RT60 for 3 CVSSP rooms related to the ground-truth (GT).

training methods such as K-fold cross-validation. We demonstrate improvements in the SSC task on the NYUv2 and NYUCAD datasets. Our previous work DBNet [9] utilises a single depth input encoded with F-TSDF for geometry representation to predict full 3D scenes. DBNet approach, enhances the model's adaptability across a variety of depth-sensing devices. In DBNet we contributed to overcoming a key challenge in SSC domain, primarily the inherent imbalance in 3D spatial distributions commonly observed in indoor scenes by introducing a re-weighting method integrated into the loss function, leveraging the K-means clustering algorithm. Although DBNet approach improved the overall mIoU score and the recognition of underrepresented classes such as chairs and tables, DBNet struggled with challenging objects, such as windows and TVs. Windows often feature reflective or transparent surfaces, while TVs share visual characteristics with other categories, such as objects, making them difficult to distinguish using depth information alone in datasets with complex scenes like NYUv2 and NYUCAD. To address this problem, we investigate the impact of learning multiple features from RGB-D input on the performance of DBNet as a case of SSC model with depth only input and propose MDBNet. We observe that incorporating RGB alongside depth features represented by F-TSDF enhances class identification both within and across object categories on NYUv2 and NYUCAD datasets. Figure 9 illustrates our previous work [9], and MDBNet SSC performance over the categories level on NYUCAD dataset. The proposed MDBNet model shows a significant improvement in overall mIoU performance compared to DBNet model. This improvement also highlights the ability of MDBNet to identify challenging object classes, such as TVs and windows, which posed significant difficulties for DBNet.

We examine various fusion strategies for integrating RGB semantic features into the proposed SSC model, including early, middle, and late fusion approaches. To ensure the robustness and generalisability of the results, K-fold cross-validation is employed. The results indicate that the model effectively learned scene semantics across different fusion methods. As shown in Table 1, the highest performance is

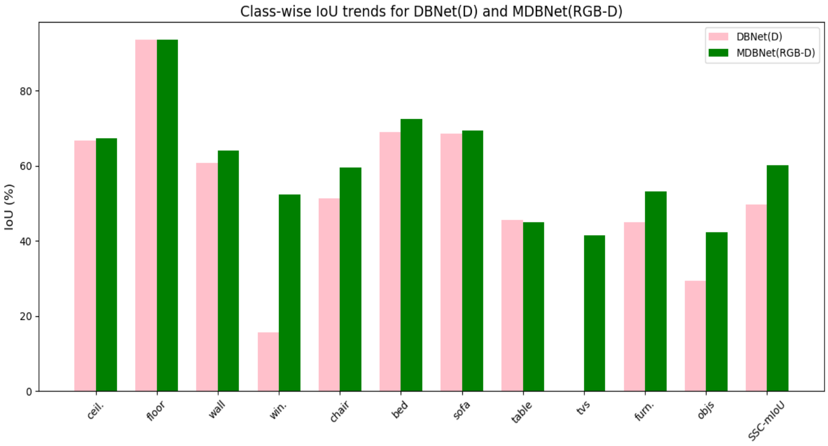


Fig. 9: IoU performance on NYUCAD dataset classes by DBNet model [9] with depth input and MDBNet model with RGB-D input.

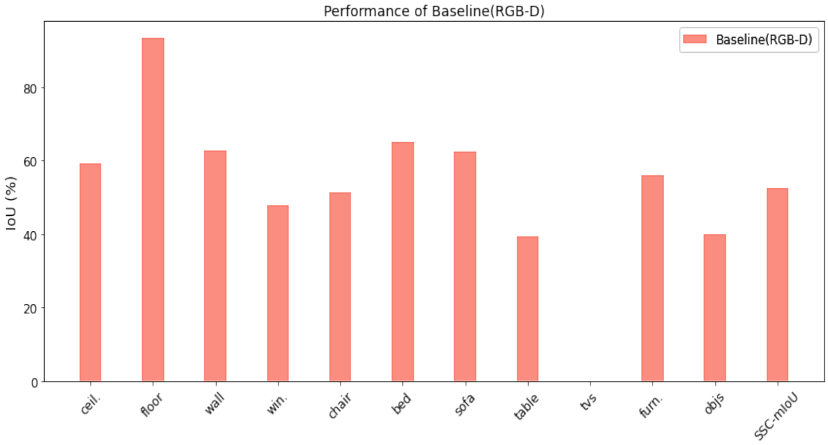


Fig. 10: The baseline architecture performance on semantics level within NYUCAD dataset.

achieved using a late fusion strategy, where learnable features are integrated through downsampling using PCR blocks to match the network's output resolution. This finding is consistent with prior SSC fusion results reported by [13]. It is important to highlight that the overall performance gains result from the combined contributions of various components within our design, rather than solely from incorporating RGB semantics. As illustrated in Figure 10, adding RGB features without integrating our proposed combined loss function leads to suboptimal results for specific categories and a lower overall mIoU score. For example, the model continues to struggle with small and rare classes, such as TVs, despite the inclusion of RGB features. This observation suggests that incorporating RGB features alone, without well-structured methodological approach, is insufficient to effectively address the challenges associated with SSC task.

However, our proposed MDBNet model is trained to predict 3D structures from perspective camera inputs, which limits its applicability in designing immersive VR spaces that require full-scene coverage. Since this research focuses on developing suitable 3D spatial modelling for VR applications, we address the limitations of perspective-based RGB-D inputs, which capture only partial views due to their restricted FOV. To overcome this, we extend the inference capabilities of the pre-trained MDBNet from perspective RGB-D to full panorama RGB-D inputs, enabling 3D SSC over full 360° surroundings. Our method leverages both RGB and depth data through a series of processing steps, as detailed in Section 3.2. We apply a spherical-to-cubic projection to the RGB data, transforming the 360° image into multiple perspective views. This transformation allow the full panorama scene to be represented as cubic faces, making it compatible with the existing perspective-based SSC model, MDBNet. Then, we perform a 3D rotation on point clouds generated from the spherical depth information to ensure proper alignment with the cubic RGB views and calculate the F-TSDF preserve the geometric structures. The processed views are then fed into the MDBNet360 model, an extension of the MDBNet architecture design specifically to handle perspective RGB-D inputs, as illustrated in Figure 3. The outputs from cubic views are fused into a unified 3D representation using a summation rule to merge overlapping regions, resulting in a comprehensive 3D model of the entire room with its full surroundings. As detailed in Section 5.2, our findings demonstrate that MDBNet360 produced more realistic scene reconstructions and improved semantic completion compared to EdgeNet360 [49], ultimately enhancing the understanding of the room's spatial structure and functional layout. We also evaluate the acoustic quality of the rendered sound within the reconstructed 3D virtual rooms generated by the proposed MDBNet360 model. Specifically, we measure the EDT and RT60 acoustic parameters, which are commonly used to characterize early reflections and late reverberations, respectively. Our results demonstrate that the 3D scenes generated by the proposed MDBNet360 produce better EDT and RT60 values compared to the SOTA models Kim19 [47] and Kim20 [49]. The block-based method in Kim19 [47] showed overestimated reverberations due to its simplified, flat surface representations [63]. Similarly, EdgeNet360's reconstructions Kim20 [4] suffer from incomplete geometry (missing chairs in the MR and KT scenes, hole in the table in the MR scene, and the absence of large portions of the central table with sofa segments in the UL scene) which compromise spatial sound propagation and increase unintended reverberation. Discontinuities and gaps in the reconstructed mesh surfaces impact the reflections of the sound waves [61, 62].

Overall, the findings confirm that the proposed SSC model not only improves the visual semantics of 3D scenes, but also enhances realism of acoustic modelling, thereby advancing the creation of immersive audio-visual VR environments. These findings illustrate the potential of our approach to bridge the gap between visual fidelity and acoustic precision, providing a foundation for a more realistic and interactive VR environment through a single 360° RGB-D input.

Fig. 11: HP Reverb G2 headset with hand controllers connected to the VR application.

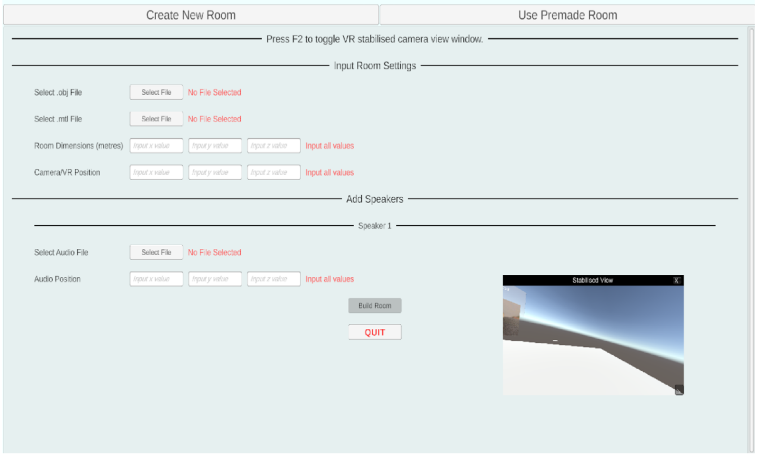
In the next section, we demonstrate a VR application of our proposed method with real time sound rendering for an immersive experience.

6 Real-time audio-visual VR rendering implementation

In this section, we demonstrate the implementation of a real-time audio-visual VR space using our proposed method on CVSSP data. The application designed based on a pipeline includes the CVSSP RGB-D inputs and DBAT material recognition model [97] for material estimation (to avoid the strong assumption of materials within the scenes), together with the SSC model (users can select either EdgeNet360 or MDBNet360) for reconstruction with full 360° surroundings. The sounds rendered in real time to provide immersive experience to the users and changes based on users movements around the sound source in the scene. For the VR demonstration, we use the HP Reverb G2 headset with controllers to manage movement and user options in the VR menu, as illustrated in Figure 11.

6.1 Unity integration

To streamline and simplify the room rendering process within Unity, a graphical user interface (GUI) is developed to enhance accessibility and usability, particularly for users with limited experience in Unity's development environment. The GUI is organised into two primary tabs: Create New Room and Use Premade Room, as illustrated in Figure 12. These tabs cater to different user needs, allowing users to either construct customized room environments or select and utilise premade ones. Premade rooms demonstrate the reconstructed 3D models with 360° surroundings in this research.



Press F2 to toggle VR stabilised camera view window.

Input Room Settings

Select .obj File | Select File | No File Selected

Select .mtl File | Select File | No File Selected

Room Dimensions (metres) | Input x value | Input y value | Input z value | Input all values

Camera/VR Position | Input x value | Input y value | Input z value | Input all values

Add Speakers

Speaker 1

Select Audio File | Select File | No File Selected

Audio Position | Input x value | Input y value | Input z value | Input all values

Fig. 12: VR application interface allows users to define and build a custom virtual environment.

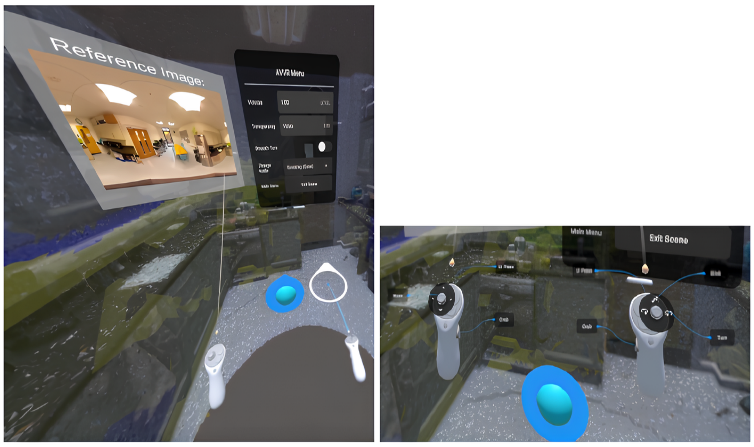


Fig. 13: VR locomotion system showing smooth movement option on the Features menu and the controllers with teleportation.

6.2 XR interaction toolkit

The VR application is built using XR interaction toolkit 3.0.3 [10] [11]. It is designed to streamline the development of immersive experiences by providing pre-built interaction components and systems for XR devices. The interaction framework can manage

[10] https://docs.unity3d.com/Packages/com.unity.xr.interaction.toolkit@3.0/manual/whats-new-3.0.html (accessed in 2024)

[11] https://medium.com/@Brian_David/xr-interaction-toolkit-reading-the-documentation-215fa825cdc6 (accessed in 2025)

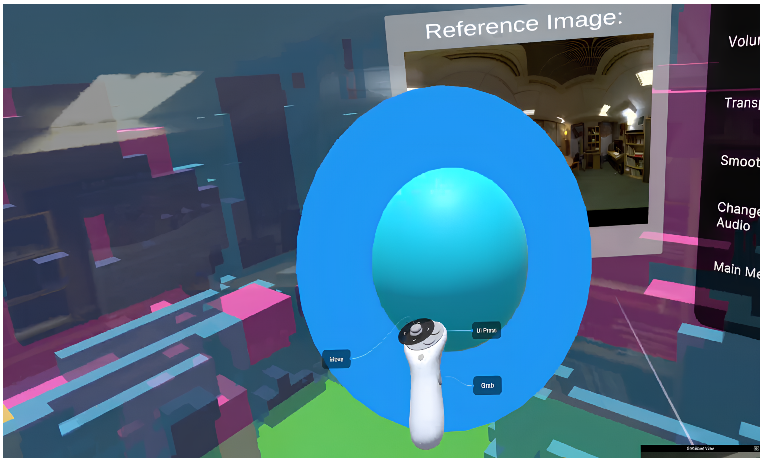


Fig. 14: Illustration of grabbing sound source sphere object (blue) within MR scene.

interactors, such as VR controllers or hands, and the objects that respond to interactions, like grabbing. Additionally, it supports interaction modes such as direct (physical contact) and ray-based (distance-based) interactions.

6.2.1 Locomotion system design

The locomotion system incorporates two primary movement modes: smooth locomotion and teleportation. Smooth locomotion is controlled via the left controller's analog stick, which allows for fluid movement through the virtual space. To mitigate potential motion sickness during movement, a dynamic FOV vignette system is implemented. This system activates during locomotion and adjusts dynamically based on movement speed to enhance user comfort during rapid movement. The teleportation is accessed through the right controller, it enables users to point to a destination on the floor plane and instantly relocate. Figure 13 illustrates the controllers and teleportation in VR.

6.2.2 Affordance system support

The XR Interaction Toolkit's affordance system enhances user interaction by providing intuitive visual feedback for interactive elements in the virtual environment. These elements respond dynamically to user proximity and interaction. We show that the user can interact with the audio source sphere in the VR space as a key example of a grabbable object implementation. Figure 14 shows the controller holding the audio source.

6.3 Features on VR menu

The VR menu system is designed to balance functionality and immersion, offering essential controls while preserving the user's sense of presence in the virtual environment.

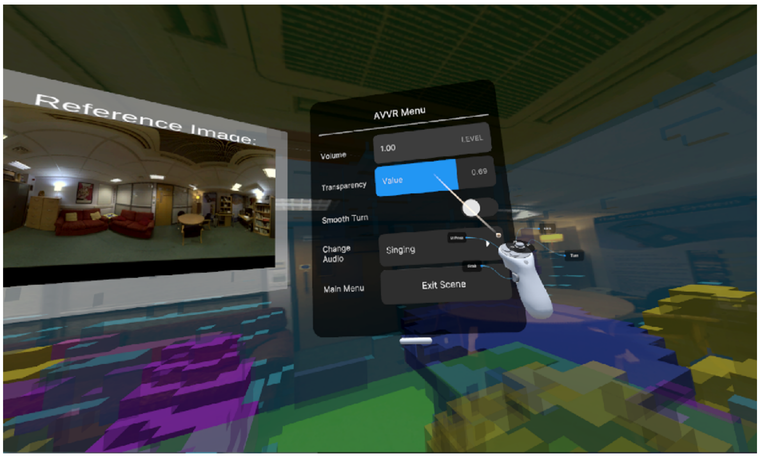


Fig. 15: VR menu showing volume and objects transparency sliders in MR scene.

6.3.1 Audio and mesh transparency controls

Real-time adjustments of both audio levels and mesh visibility are managed through intuitive slider controls as illustrated in Figure 15. The volume slider allows for precise tuning with visual feedback, while the mesh transparency slider enables users to seamlessly transition between the reconstructed geometry and the original reference image. Users can adjust audio volume through VR controllers and modify the transparency of reconstructed meshes in real-time. Users can adjust the sound volume of the audio source from the VR menu, and we found that the sound level dynamically varies based on the user's distance from the source, resulting in a more realistic audio effect. Furthermore, the audio system supports multiple options, including music and speech samples at varying volumes as illustrated in Figure 16.

On the other side, we implement RGB textures and LiDAR scans as visual references. For example, for MR and UL scenes, RGB textures are used, whereas LiDAR scans are used for the KT scene. The LiDAR scan enhances the experience with detailed textures of the scene, which increases immersion and presence within the VR space. Also, it serves as a spatial reference for the scene components. Figure 17 illustrate the LiDAR integration to KT scene while keeping the reconstructed 3D model transparent. However, these visual references help users better understand the scene components and enhance the perceived realism of the reconstructed environments.

6.3.2 Movement and spacial interactions

To accommodate user comfort, the application allows toggling between snap turning and smooth turning. Snap turning applies fixed-angle rotations and is often preferred to reduce motion sickness caused by continuous camera movement. The audio source maintains proper spatial audio properties with position-based attenuation, enhancing the overall immersion of the experience. The application allows users to interact with and manipulate sound sources by grabbing and repositioning them within the scenes.

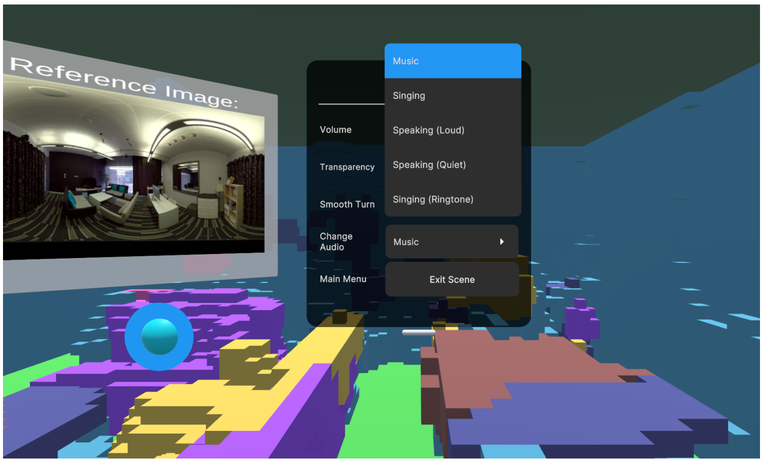


Fig. 16: VR menu showing the audio options in UL scene.

Through testing, we observed that the spatial audio system provides realistic reverberations, thus enhancing immersion. The application also adjusts sound propagation based on sound source location and user movements within the scene. For example, when a sound source is placed in an occluded region, such as at the center of an object's voxel mesh (e.g., clipped inside a wall), the perceived volume is noticeably reduced or muted if fully occluded. The sound reduced when placed under the table in the scene (e.g., place the audio source under the table in the KT scene) providing a realistic experience.

The proposed application serves as a demonstration of an immersive experience integrating spatial audio and visual cues within a VR environment. Users can move freely within the virtual space and interact with sound source. Future improvements can include expanding interactive elements to enhance user engagement within the scene.

A demonstration video of the VR space generated using 3D scenes by MDBNet360, showcasing the application's functionality, is available at: https://github.com/MonaIA1/Repo360.

6.4 Subjective evaluation

A user evaluation was conducted to assess the overall effectiveness of the proposed VR application in delivering an immersive audio–visual experience. The evaluation was carried out using a pipeline that combined both the EdgeNet360 and MDBNet360 frameworks, allowing users to freely select and explore different virtual environments. In this evaluation, we focus on assessing the overall user experience, with a particular emphasis on the audio and visual realism of the scenes reconstructed primarily using EdgeNet360 to provide a bottom-line performance evaluation of the perceptual quality of the system's audio-visual rendering.

To ensure consistency and participant well-being, each volunteer was briefed on the study's purpose, objectives, and procedures, and completed a health screening

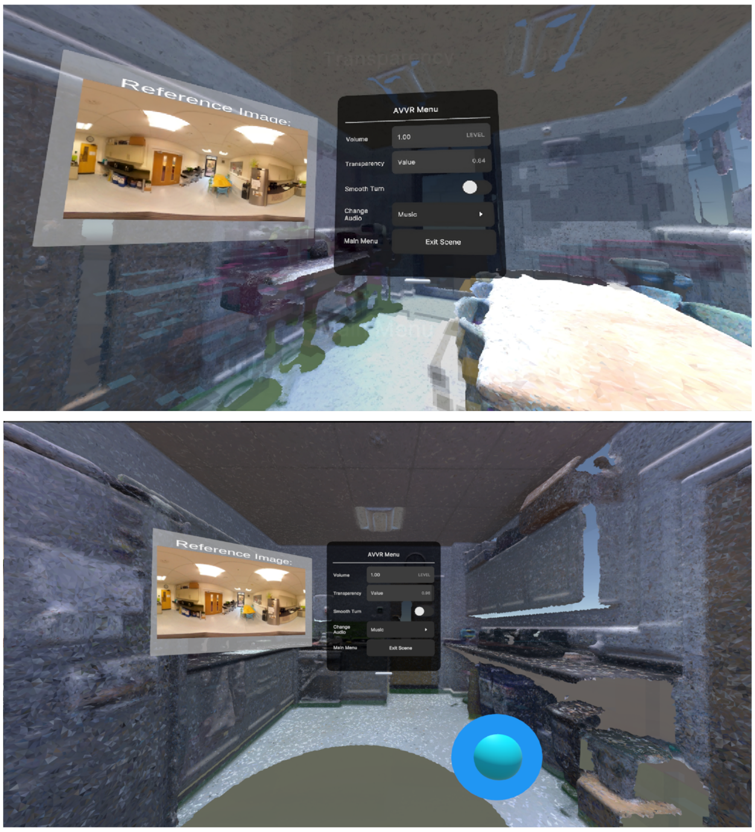
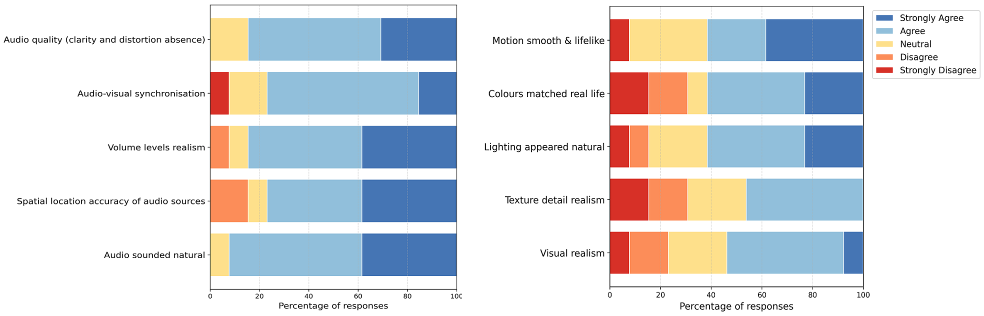


Fig. 17: Illustration of the KT scene with an overlaid LiDAR scan with two different view points.

questionnaire to identify potential risks such as motion sickness or pre-existing medical conditions that might be aggravated by VR usage [98]. Participants who passed the screening were provided with an information sheet outlining their rights and signed a consent form before taking part in the study. The study was approved by the authors' local institutional ethics committee under reference number ERGO/FEPS/99833.

During the evaluation, each participant independently explored a pre-loaded VR environment, ensuring consistent audio and visual fidelity across all sessions. The evaluation focused on spatial accuracy of audio-visual environment, and the overall user experience in the virtual space. Minimal guidance was provided to allow participants to naturally assess the system's usability and intuitive operation. Following the exploration phase, participants completed a structured questionnaire containing statements addressing these aspects, rated on a five-point Likert scale (Strongly Disagree, Disagree, Neutral, Agree, Strongly Agree) [99]. To complement the quantitative feedback, participants were also encouraged to provide open-ended comments and suggestions, offering qualitative insights into their experiences.



(a) Audio Quality

(b) Visual Quality

Fig. 18: Responses on evaluating audio and visual aspects in the VR space.

Responses from 13 completed questionnaires were analysed and categorized into three key themes: audio quality, visual quality, and overall user experience. The findings revealed that participants generally rated the system as immersive and realistic, with particularly strong satisfaction in spatial audio quality and environmental realism. Some noted minor visual limitations related to texture detail and colour consistency, identifying potential areas for refinement. Overall, the evaluation confirmed the system's ability to provide an engaging and comfortable audio–visual experience for VR users.

6.4.1 Audio-visual 3D spatial evaluation

The evaluation focused on three key aspects: (1) the realism and spatial accuracy of audio (e.g., reverberation, occlusion, and spatial positioning), (2) the visual fidelity of the environment (e.g., lighting, colour, and surface detail), and (3) the overall user experience in the virtual space. Figure 18 summarises participants' ratings for audio (left) and visual (right) aspects using a five-point Likert scale. Figure 19 presents the participants' evaluation of their overall experience, also using a five-point Likert scale.

Audio quality. Responses were strongly positive across all items as shown in '(a)' within Figure 18. Perceived audio quality achieved 85% positive agreement (Agree/Strongly Agree) in terms of clarity and absence of distortion, with no negative responses. Audio–visual synchronisation was rated positively by 77% of participants. Approximately 85% reported that the volume levels of different sounds appeared realistic and that changes in distance between the audio source and listener accurately mimicked real-world acoustic behaviour. The spatial localisation accuracy of audio sources reached 77%, while 92% of participants agreed that the overall audio sounded natural and realistic, confirming the high audio fidelity of the system. These findings validate the effectiveness of the real-time spatial audio rendering integrated with the 3D visual model, which includes occlusion, distance attenuation, and reverberation effects.

Visual quality. Participants generally provided positive feedback regarding the visual representation of the VR scene as illustrated in '(b)' within Figure 18. About 62% agreed that the motion of objects appeared lifelike, while 31% gave neutral responses and one participant disagreed. This may be attributed to the fact that the only movable element in the scene was the audio source, which might have limited participants' ability to assess motion realism. Similarly, 62% of participants agreed that the scene colours appeared realistic, although some found the colour tones less natural. To address this, additional attention should be given to ensuring that colours are well-saturated and consistent with real-world lighting by applying cube-map colour grading and tone adjustments. The naturalness of the lighting was also positively rated by 62% of the participants, with only a few negative responses, suggesting potential areas for improvement. Adjusting colour temperature according to the input image rather than using a uniform illumination across all scenes could help achieve a more realistic and immersive appearance. Texture detail (e.g., surfaces and objects) received mixed feedback, with 46% positive and 31% negative responses. This indicates a need for higher-resolution cube maps and the incorporation of texture maps to improve surface detail. 54% of the participants agreed that the visual representation of the scene appeared realistic, which confirms the satisfactory representations in visual VR.

Overall user experience. The overall user experience evaluation revealed mixed responses regarding realism and presence but consistently positive feedback on immersion as shown in Figure 19. All participants agreed that the combined audio–visual experience felt immersive, confirming the strong objective level of the sensory fidelity of the VR application through coherent spatial audio and stable visual rendering. However, the relatively high number of neutral ratings for realism and 31% negative ones observed for the sense of presence indicates that the environment's visual fidelity did not fully meet participants' expectations. This suggests that while the system effectively delivers sensory immersion, it could further benefit from improvements to visual components such as texture, lighting, and interactive objects previously identified as visual limitations. Enhancing these elements particularly through higher-resolution texture mapping, more dynamic lighting effects, and additional moving objects, could strengthen the perception of realism and sense of presence.

Overall, the findings confirm that the system successfully conveys immersion through integrated audio–visual design, but that refining the visual components remains essential for achieving stronger perceptual realism and presence. Incorporating the user feedback from this evaluation provides a clear direction for improving future iterations of the VR application.

Participants' qualitative feedback. Over half of the participants provided additional comments elaborating on their experiences during the VR evaluation, offering insights that complemented the questionnaire results. Several participants praised the realism and immersion provided by the audio component. One noted that the "sound immersion is quite accurate," expressing surprise at the precision of spatial audio effects. Another commented that "audio can express the effect of Doppler," acknowledging that the system successfully reproduced real-world acoustic behaviour through wave-based simulation. A further participant observed that "the audio changes depending on where it is placed in the scene," referring to the perceived variation in reverberation and damping caused by surrounding materials and nearby objects. Collectively, these comments underscore participants' strong satisfaction with the realism achieved through real-time audio simulation.

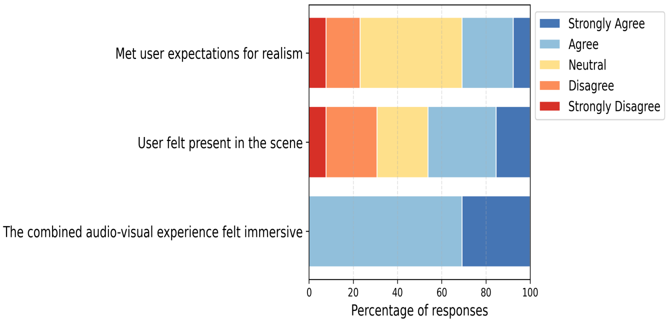


Fig. 19: Overall user-experience evaluation.

Feedback on the visual aspects, however, indicated areas for improvement. One participant mentioned that “the mesh and image in the scene could be improved,” referring to the original 360° image used to generate the room. This limitation was linked to the relatively low resolution of the depth input and RGB clarity, which contributed to reduced texture and mesh quality in reconstruction. Another participant noted that “the ceilings were a bit rounded on the edges and could be improved,” pointing to geometric inaccuracies that diminished perceived realism. Such observations emphasise the importance of refining the underlying depth estimation and geometry reconstruction algorithms to enhance structural fidelity and visual coherence.

Together, these qualitative insights highlight that the system’s audio design was consistently praised for its realism and immersion, while visual fidelity particularly in geometric detail was recognised as the primary area for future enhancement.

7 Conclusion and future work

This work addresses the complex and largely underexplored challenge of creating a 3D representation of real-world indoor spaces that integrates both visually accurate geometry and acoustically plausible spatial audio from a single 360° RGB-D input. We introduce MDBNet, a 3D SSC model trained on perspective RGB-D data, and extended its capabilities to full panorama inputs through MDBNet360. Our method leverages spherical-to-cubic projection for RGB data and applies 3D rotation to point clouds derived from depth, enabling the construction of detailed 3D models that capture full 360° spatial context. While effective, the cubic projection introduces distortions near cube face edges and may not generalise well to diverse room shapes. Future work could explore more advanced projection methods to address these limitations. Furthermore, in our MDBNet model, we examined the use of the Tanh activation function on identity features within the proposed ITRM block. While this approach demonstrated a stabilizing effect, it merits further investigation. Future research could explore the role of Tanh activation in cross-modal architectures and TSDF-based inputs, potentially offering deeper insights into optimizing non-linear transformations for SSC. In our MDBNet model, we evaluated performance using a combined loss in which the 3D component is adapted from our previous work, DBNet. While this formulation proved effective, future research could compare it with other imbalance-handling strategies (such as focal, Asymmetric Loss (ASL) ,or LDAM losses [100–102]) to further assess its relative advantages.

Another challenge lies in the lack of datasets that jointly provide RIRs and 360° RGB-D data. Although datasets like Matterport3D [103] and 2D-3D-S [104] offer high-quality 3D reconstructions, they lack acoustic annotations and often require post-processing to repair surface discontinuities. To bridge this gap, we employe the CVSSP dataset, which uniquely combines measured RIRs with 360° RGB-D scenes. Our approach yielded promising results, as detailed in Section 5. However, generalisation remains limited. For instance, sound artifacts observed in one scene may stem from inherent limitations in the Steam Audio plug-in's spatial modelling. Rather than detracting from our results, this highlights the importance of further research in spatial audio integration. Future work should focus on expanding multimodal datasets with more diverse indoor scenes, accurate 3D annotations, and corresponding RIRs. While collecting such data is non-trivial due to high degree of occlusion, the diversity of objects in indoor scene, and the time and cost involved in acquiring RIRs in real space, it is essential for building truly immersive and generalisable VR experiences. This research contributes through the horizontal integration of AI and VR, bridging the gap between visual and auditory realism in 3D virtual spaces.

Funding

This work was supported by Electronics and Telecommunications Research Institute (ETRI) grant funded by the Korean government (24ZC1200, Research on hyper-realistic interaction technology for five senses and emotional experience).

Consent Statement

This study involved human participants for subjective evaluation and was approved by the authors' local institutional ethics committee under reference number ERGO/FEPS/99833.

References

[1] Mandal, S.: Brief introduction of virtual reality & its challenges. International Journal of Scientific & Engineering Research **4**(4), 304–309 (2013)

[2] Stecker, G.C., Moore, T.M., Folkerts, M., Zotkin, D., Duraiswami, R.: Toward objective measures of auditory co-immersion in virtual and augmented reality. In: Audio Engineering Society Conference: 2018 AES International Conference on Audio for Virtual and Augmented Reality (2018). Audio Engineering Society

[3] Privitera, A.G., Fontana, F., Geronazzo, M.: The role of audio in immersive storytelling: a systematic review in cultural heritage. Multimedia Tools and Applications, 1–39 (2024)

[4] Kim, H., Remaggi, L., Jackson, P.J.B., Hilton, A.: Immersive virtual reality audio rendering adapted to the listener and the room. In: Magnor, M., Sorkine-Hornung, A. (eds.) Real VR–Immersive Digital Reality: How to Import the Real World Into Head-Mounted Immersive Displays, pp. 293–318. Springer, Cham (2020)

[5] Berkman, M.I.: History of virtual reality. In: Lee, N. (ed.) Encyclopedia of Computer Graphics and Games, pp. 873–881. Springer, Cham (2024)

[6] Song, S., Yu, F., Zeng, A., Chang, A.X., Savva, M., Funkhouser, T.: Semantic scene completion from a single depth image. In: CVPR, pp. 1746–1754 (2017)

[7] Pan, Y., Xie, F., Zhao, H.: Understanding the challenges when 3d semantic segmentation faces class imbalanced and ood data. IEEE Transactions on Intelligent Transportation Systems **24**(7), 6955–6970 (2023)

[8] Kon, H., Koike, H.: Deep neural networks for cross-modal estimations of acoustic reverberation characteristics from two-dimensional images. In: Audio Engineering Society Convention 144 (2018)

[9] Alawadh, M., Niranjan, M., Kim, H.: 3d semantic scene completion from a depth map with unsupervised learning for semantics prioritisation. In: 2024 IEEE International Conference on Image Processing (ICIP), pp. 3348–3354 (2024). IEEE

[10] Van Damme, S., Vega, M.T., De Turck, F.: Human-centric quality management of immersive multimedia applications. In: 2020 6th IEEE Conference on Network Softwarization (NetSoft), pp. 57–64 (2020). IEEE

[11] Doolani, S., Wessels, C., Kanal, V., Sevastopoulos, C., Jaiswal, A., Nambiappan, H., Makedon, F.: A review of extended reality (xr) technologies for manufacturing training. Technologies **8**(4), 77 (2020)

[12] Partarakis, N., Zabulis, X.: A review of immersive technologies, knowledge representation, and ai for human-centered digital experiences. Electronics **13**(2), 269 (2024)

[13] Roldao, L., De Charette, R., Verroust-Blondet, A.: 3d semantic scene completion: a survey. IJCV, 1–28 (2022)

[14] Garbade, M., Chen, Y.-T., Sawatzky, J., Gall, J.: Two stream 3d semantic scene completion. In: CVPRW, pp. 0–0 (2019)

[15] Zhang, L., Wang, L., Zhang, X., Shen, P., Bennamoun, M., Zhu, G., Shah, S.A.A., Song, J.: Semantic scene completion with dense crf from a single depth image. Neurocomputing **318**, 182–195 (2018)

[16] Zhang, J., Zhao, H., Yao, A., Chen, Y., Zhang, L., Liao, H.: Efficient semantic scene completion network with spatial group convolution. In: ECCV, pp. 733–749 (2018)

[17] Zhang, P., Liu, W., Lei, Y., Lu, H., Yang, X.: Cascaded context pyramid for full-resolution 3d semantic scene completion. In: ICCV, pp. 7801–7810 (2019)

[18] Dourado, A., De Campos, T.E., Kim, H., Hilton, A.: Edgenet: Semantic scene completion from a single rgb-d image. In: ICPR, pp. 503–510 (2021)

[19] Li, J., Song, Q., Yan, X., Chen, Y., Huang, R.: From front to rear: 3d semantic scene completion through planar convolution and attention-based network. IEEE TMM (2023)

[20] Liu, S., Hu, Y., Zeng, Y., Tang, Q., Jin, B., Han, Y., Li, X.: See and think: Disentangling semantic scene completion **31** (2018)

[21] Li, S., Zou, C., Li, Y., Zhao, X., Gao, Y.: Attention-based multi-modal fusion network for semantic scene completion. In: AAAI, pp. 11402–11409 (2020)

[22] Li, J., Liu, Y., Gong, D., Shi, Q., Yuan, X., Zhao, C., Reid, I.: Rgbd based dimensional decomposition residual network for 3d semantic scene completion. In: CVPR, pp. 7693–7702 (2019)

[23] Li, J., Han, K., Wang, P., Liu, Y., Yuan, X.: Anisotropic convolutional networks for 3d semantic scene completion. In: CVPR, pp. 3351–3359 (2020)

[24] Zhong, M., Zeng, G.: Semantic point completion network for 3d semantic scene completion. In: De Giacomo, G., Català, A., Montalvo, B., Rossi, F. (eds.) European Conference on Artificial Intelligence, pp. 2824–2831. IOS Press, Amsterdam (2020)

[25] Cao, A.-Q., Charette, R.: Monoscene: Monocular 3d semantic scene completion. In: CVPR, pp. 3991–4001 (2022)

[26] Wang, X., Feng, W., Wan, L.: Multi-modal fusion architecture search for camera-based semantic scene completion. Expert Systems with Applications **243**, 122885 (2024)

[27] Yao, J., Li, C., Sun, K., Cai, Y., Li, H., Ouyang, W., Li, H.: Ndc-scene: Boost monocular 3d semantic scene completion in normalized device coordinates space. In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 9421–9431 (2023). IEEE Computer Society

[28] Li, J., Liu, Y., Yuan, X., Zhao, C., Siegwart, R., Reid, I., Cadena, C.: Depth based semantic scene completion with position importance aware loss. IEEE Robotics and Automation Letters **5**(1), 219–226 (2019)

[29] Chen, X., Lin, K.-Y., Qian, C., Zeng, G., Li, H.: 3d sketch-aware semantic scene completion via semi-supervised structure prior. In: CVPR, pp. 4193–4202 (2020)

[30] Cai, Y., Chen, X., Zhang, C., Lin, K.-Y., Wang, X., Li, H.: Semantic scene completion via integrating instances and scene in-the-loop. In: CVPR, pp. 324–333 (2021)

[31] Wang, X., Lin, D., Wan, L.: Ffnet: Frequency fusion network for semantic scene completion. In: AAAI, pp. 2550–2557 (2022)

[32] Dourado, A., Guth, F., Campos, T.: Data augmented 3d semantic scene completion with 2d segmentation priors. In: IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 3781–3790 (2022)

[33] Wang, F., Zhang, D., Zhang, H., Tang, J., Sun, Q.: Semantic scene completion with cleaner self. In: CVPR, pp. 867–877 (2023)

[34] Li, J., Ding, L., Huang, R.: Imenet: Joint 3d semantic scene completion and 2d semantic segmentation through iterative mutual enhancement. In: IJCAI (2021)

[35] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR, pp. 770–778 (2016)

[36] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., *et al.*: Imagenet large scale visual recognition challenge. IJCV **115**(3), 211–252 (2015)

[37] Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: ECCV, pp. 801–818 (2018)

[38] ADE20K dataset. https://tinyurl.com/ADE20K. [Online; accessed 2023-01-17]

[39] Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P.: Segformer: Simple and efficient design for semantic segmentation with transformers, pp. 12077–12090 (2021)

[40] Tang, J., Chen, X., Wang, J., Zeng, G.: Not all voxels are equal: Semantic scene completion from the point-voxel perspective. In: AAAI, pp. 2352–2360 (2022)

[41] Lee, S., Chung, J., Huh, J., Lee, K.M.: Odgs: 3d scene reconstruction from omnidirectional images with 3d gaussian splattings. Advances in Neural Information Processing Systems **37**, 57050–57075 (2024)

[42] Han, H., Liang, Y., Zhou, Y., Wang, W., J. Rojas-Muñoz, E., Li, X.: Aurora: Automated unleash of 3d room outlines for vr applications. In: Proceedings of the 19th ACM SIGGRAPH International Conference on Virtual-Reality

Continuum and Its Applications in Industry, pp. 1–8 (2024)

[43] Li, T., Zhang, Z., Wang, Y., Cui, Y., Li, Y., Zhou, D., Yin, B., Yang, X.: Self-supervised indoor scene point cloud completion from a single panorama. The Visual Computer, 1–15 (2024)

[44] Yang, S.-T., Wang, F.-E., Peng, C.-H., Wonka, P., Sun, M., Chu, H.-K.: Dula-net: A dual-projection network for estimating room layouts from a single rgb panorama. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3363–3372 (2019)

[45] Meng, M., Zhou, Y., Zuo, D., Li, Z., Zhou, Z.: Structure recovery from single omnidirectional image with distortion-aware learning. Journal of King Saud University-Computer and Information Sciences **36**(7), 102151 (2024)

[46] Li, M., Meng, M., Zhou, Z.: Repf-net: Distortion-aware re-projection fusion network for object detection in panorama image. In: Proceedings of the Asian Conference on Computer Vision, pp. 74–89 (2022)

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

[48] Badrinarayanan, V., Kendall, A., Cipolla, R.: Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE transactions on pattern analysis and machine intelligence **39**(12), 2481–2495 (2017)

[49] Kim, H., Remaggi, L., Dourado, A., Campos, T.d., Jackson, P.J., Hilton, A.: Immersive audio-visual scene reproduction using semantic scene reconstruction from 360 cameras. Virtual Reality **26**(3), 823–838 (2022)

[50] Remaggi, L., Jackson, P., Coleman, P.: Estimation of room reflection parameters for a reverberant spatial audio object. In: Audio Engineering Society Convention 138 (2015)

[51] Politis, A., Tervo, S., Lokki, T., Pulkki, V.: Parametric multidirectional decomposition of microphone recordings for broadband high-order ambisonic encoding. In: Audio Engineering Society Convention 144 (2018). Audio Engineering Society

[52] Baran, M.-V., King, R., Woszczyk, W.: A general overview of methods for generating room impulse responses. The Journal of the Acoustical Society of America **155**(3_Supplement), 282–282 (2024)

[53] Raghuvanshi, N., Snyder, J., Mehra, R., Lin, M., Govindaraju, N.: Precomputed wave simulation for real-time sound propagation of dynamic sources in complex scenes. In: ACM SIGGRAPH 2010 Papers, pp. 1–11 (2010)

[54] Lentz, T., Schröder, D., Vorländer, M., Assenmacher, I.: Virtual reality system with integrated sound field simulation and reproduction. EURASIP journal on advances in signal processing **2007**, 1–19 (2007)

[55] Taylor, M., Chandak, A., Mo, Q., Lauterbach, C., Schissler, C., Manocha, D.: Guided multiview ray tracing for fast auralization. IEEE transactions on visualization and computer graphics **18**(11), 1797–1810 (2012)

[56] Chen, M., Su, K., Shlizerman, E.: Be everywhere-hear everything (bee): Audio scene reconstruction by sparse audio-visual samples. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 7853–7862 (2023)

[57] Liang, S., Huang, C., Tian, Y., Kumar, A., Xu, C.: Neural acoustic context field: Rendering realistic room impulse response with neural fields. arXiv preprint arXiv:2309.15977 (2023)

[58] Majumder, S., Chen, C., Al-Halah, Z., Grauman, K.: Few-shot audio-visual learning of environment acoustics. Advances in Neural Information Processing Systems **35**, 2522–2536 (2022)

[59] Ratnarajah, A., Ghosh, S., Kumar, S., Chiniya, P., Manocha, D.: Av-rir: Audio-visual room impulse response estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 27164–27175 (2024)

[60] Singh, N., Mentch, J., Ng, J., Beveridge, M., Drori, I.: Image2reverb: Cross-modal reverb impulse response synthesis. In: ICCV, pp. 286–295 (2021)

[61] Wang, R., Zhang, Y., Jia, B.: Research on the influence of object surface discontinuity on target acoustic scattering characteristics. In: 2021 6th International Conference on Communication, Image and Signal Processing (CCISP), pp. 345–349 (2021). IEEE

[62] Torres, R., Rycker, N., Kleiner, M.: Edge diffraction and surface scattering in concert halls: physical and perceptual aspects. Journal of Temporal Design in Architecture and the Environment **4**(1), 52–58 (2004)

[63] Shtrepi, L.: Investigation on the diffusive surface modeling detail in geometrical acoustics based simulations. The Journal of the Acoustical Society of America **145**(3), 215–221 (2019)

[64] Isar, C.: A glance into virtual reality development using unity. Informatica Economica **22**(3), 14–22 (2018)

[65] Anil, Ç.: Modern workflows for procedural audio at the intersection of gaming

and music performance in virtual reality. In: Audio Engineering Society Conference: AES 2024 International Audio for Games Conference (2024). Audio Engineering Society

[66] Sabir, A., Hussain, R., Pedro, A., Soltani, M., Lee, D., Park, C., Pyeon, J.-H.: Synthetic data generation with unity 3d and unreal engine for construction hazard scenarios: A comparative analysis

[67] Ciekanowska, A., Kiszczak-Gliński, A., Dziedzic, K.: Vr space such as. Journal of Computer Sciences Institute **20**, 247–253 (2021)

[68] Wolf, M., Trentsios, P., Kubatzki, N., Urbanietz, C., Enzner, G.: Implementing continuous-azimuth binaural sound in unity 3d. In: 2020 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW), pp. 384–389 (2020). IEEE

[69] Røsvik, P.M.: Creating a virtual reality orchestral concert experience with 3d audio. Master's thesis, The University of Bergen (2024)

[70] Barron, M.: Interpretation of early decay times in concert auditoria. Acta Acustica united with Acustica **81**(4), 320–331 (1995)

[71] Rungta, A., Rust, S., Morales, N., Klatzky, R., Lin, M., Manocha, D.: Psychoacoustic characterization of propagation effects in virtual environments. ACM Transactions on Applied Perception (TAP) **13**(4), 1–18 (2016)

[72] Bradley, J.S.: Review of objective room acoustics measures and future needs. Applied Acoustics **72**(10), 713–720 (2011)

[73] Dunn, F., Hartmann, W., Campbell, D., Fletcher, N.H.: Springer Handbook of Acoustics. Springer, New York (2015)

[74] Park, J.J., Florence, P.R., Straub, J., Newcombe, R.A., Lovegrove, S.: Deepsdf: Learning continuous signed distance functions for shape representation. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 165–174 (2019)

[75] Weder, S., Schönberger, J.L., Pollefeys, M., Oswald, M.R.: Neuralfusion: Online depth fusion in latent space. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 3161–3171 (2020)

[76] Silberman, N., Hoiem, D., Kohli, P., Fergus, R.: Indoor segmentation and support inference from rgbd images. In: ECCV, pp. 746–760 (2012)

[77] Firman, M., Mac Aodha, O., Julier, S., Brostow, G.J.: Structured prediction of unobserved voxels from a single depth image. In: CVPR, pp. 5431–5440 (2016)

[78] Kim, H., Hilton, A.: Block world reconstruction from spherical stereo image

pairs. Computer Vision and Image Understanding **139**, 104–121 (2015)

[79] Pi, H., Tian, S., Lu, M., Liu, J., Guo, Y., Zhang, S.: A comprehensive comparison of projections in omnidirectional super-resolution. In: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1–5 (2023). IEEE

[80] Wang, Y.: Projection methods for 360-degree video. Frontiers in Computing and Intelligent Systems (2024)

[81] Kittler, J., Hatef, M., Duin, R.P., Matas, J.: On combining classifiers. IEEE transactions on pattern analysis and machine intelligence **20**(3), 226–239 (1998)

[82] Farina, A.: Advancements in impulse response measurements by sine sweeps. In: Audio Engineering Society Convention 122 (2007). Audio Engineering Society

[83] Farina, A.: Simultaneous measurement of impulse response and distortion with a swept-sine technique. In: Audio Engineering Society Convention 108 (2000). Audio Engineering Society

[84] Močnik, M.: pyrirtool: A Python Tool for Room Impulse Response (RIR) Processing. https://github.com/maj4e/pyrirtool. Accessed: 2024-10-02 (2023)

[85] International Organization for Standardization: ISO 3382-1:2009: Acoustics – Measurement of Room Acoustic Parameters – Part 1: Performance Spaces. https://www.iso.org/standard/40979.html (2009)

[86] IoSR: IoSR Matlab Toolbox. https://github.com/IoSR-Surrey/MatlabToolbox/tree/master. Accessed: 2024-10-02 (2024)

[87] Meng, Z., Zhao, F., He, M.: The just noticeable difference of noise length and reverberation perception. In: 2006 International Symposium on Communications and Information Technologies, pp. 418–421 (2006). IEEE

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

[89] Yao, Y., Mihalcea, R.: Modality-specific learning rates for effective multimodal additive late-fusion. In: The Association for Computational Linguistics (ACL), pp. 1824–1834 (2022)

[90] NVIDIA: SegFormer B5 Finetuned ADE 640x640. http://tinyurl.com/segformerb5. Accessed: 2024-02-06 (2024)

[91] Moradi, R., Berangi, R., Minaei, B.: A survey of regularization strategies for deep models. Artificial Intelligence Review **53**(6), 3947–3986 (2020)

[92] Stone, M.: Cross-validatory choice and assessment of statistical predictions. Journal of the royal statistical society: Series B (Methodological) **36**(2), 111–133 (1974)

[93] Wong, T.-T.: Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation. Pattern Recognition **48**(9), 2839–2846 (2015)

[94] Rodriguez, J.D., Perez, A., Lozano, J.A.: Sensitivity analysis of k-fold cross validation in prediction error estimation. IEEE transactions on pattern analysis and machine intelligence **32**(3), 569–575 (2009)

[95] Liu, X., Xie, H., Zhang, S., Yao, H., Ji, R., Nie, L., Tao, D.: 2d semantic-guided semantic scene completion. International Journal of Computer Vision, 1–20 (2024)

[96] Wang, F., Sun, Q., Zhang, D., Tang, J.: Unleashing network potentials for semantic scene completion. In: CVPR, pp. 10314–10323 (2024)

[97] Heng, Y., Dasmahapatra, S., Kim, H.: Dbat: Dynamic backward attention transformer for material segmentation with cross-resolution patches. arXiv preprint arXiv:2305.03919 (2023)

[98] Chang, E., Kim, H.T., Yoo, B.: Virtual reality sickness: a review of causes and measurements. International Journal of Human–Computer Interaction **36**(17), 1658–1682 (2020)

[99] Joshi, A., Kale, S., Chandel, S., Pal, D.K.: Likert scale: Explored and explained. British journal of applied science & technology **7**(4), 396 (2015)

[100] Lin, T.-Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: ICCV, pp. 2980–2988 (2017)

[101] Ridnik, T., Ben-Baruch, E., Zamir, N., Noy, A., Friedman, I., Protter, M., Zelnik-Manor, L.: Asymmetric loss for multi-label classification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 82–91 (2021)

[102] Cao, K., Wei, C., Gaidon, A., Arechiga, N., Ma, T.: Learning imbalanced datasets with label-distribution-aware margin loss. Advances in neural information processing systems **32** (2019)

[103] 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. arXiv preprint arXiv:1709.06158 (2017)

[104] Armeni, I., Sax, S., Zamir, A.R., Savarese, S.: Joint 2d-3d-semantic data for indoor scene understanding. arXiv preprint arXiv:1702.01105 (2017)