Virtual Reality (2026) 30:55
https://doi.org/10.1007/s10055-026-01312-7

ORIGINAL ARTICLE

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]

Received: 13 July 2025 / Accepted: 5 January 2026 / Published online: 6 February 2026

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 (Mandal 2013). Both visual and synchronised spatial audio are essential for creating truly immersive environment experiences (Stecker et al. 2018; Privitera et al. 2024; Kim et al. 2020). 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 (Berkman 2024). 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

✉ Mona Alawadh
malawadh@imamu.edu.sa

✉ Hansung Kim
h.kim@soton.ac.uk

Atiyeh Alinaghi
a.alinaghi@soton.ac.uk

Mahesan Niranjan
mn@ecs.soton.ac.uk

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

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

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 (Song et al. 2017). 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 (Pan et al. 2023). 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 (Kon and Koike 2018). 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.

In this research, as illustrated in Fig. 1 we extend our previous model for SSC using depth-only input (Alawadh et al. 2024), which we refer to as 'DBNet'. We propose a new hybrid architecture, MDBNet, which features a dual-head

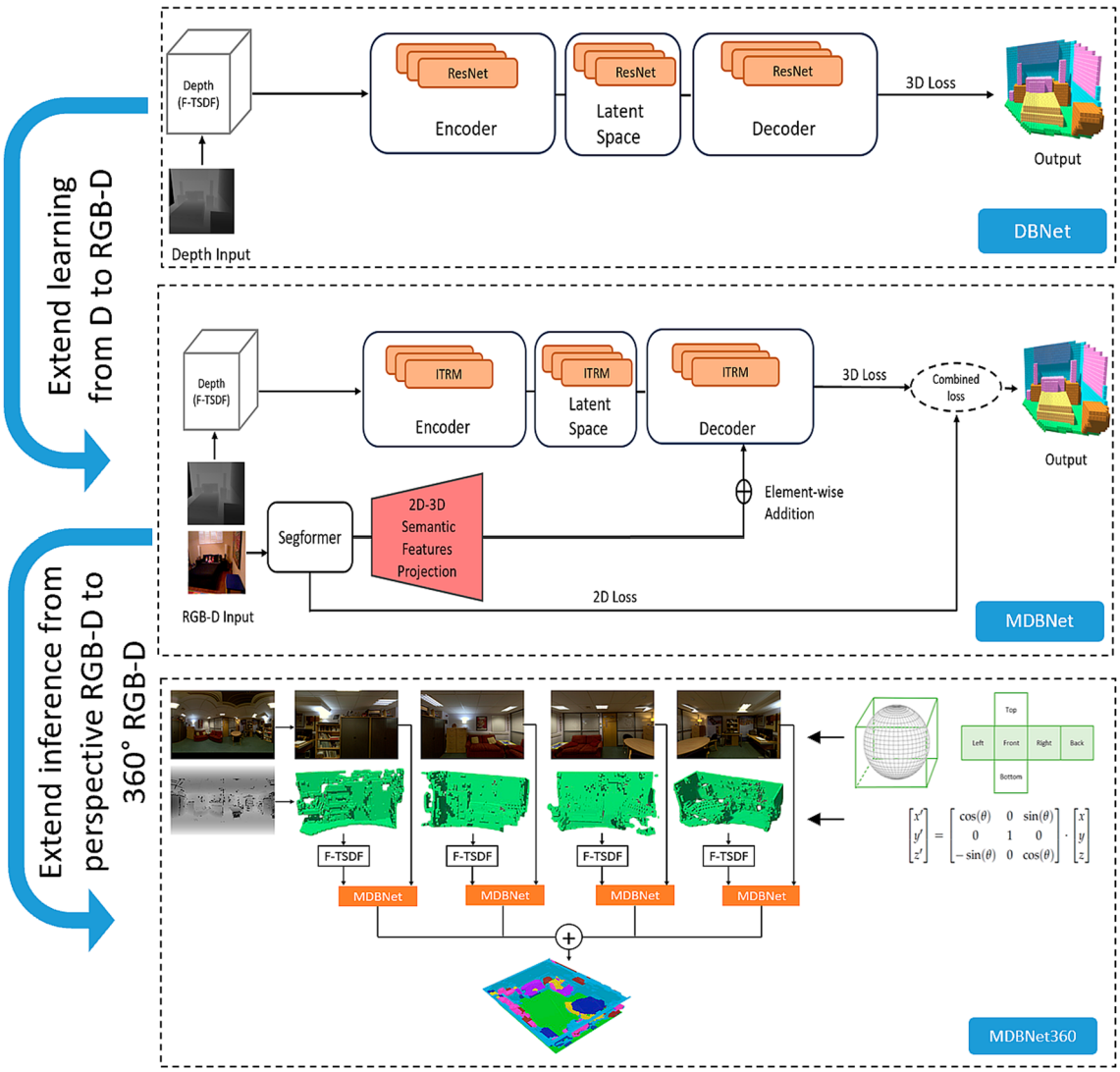


Fig. 1 System stages improvement to construct 3D space starting by DBNet with depth only input then MDBNet with RGB-D, ending with MDBNet360 with RGB-D input

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 (Van Damme et al. 2020), and contributes to revolutionising experiential learning paradigms (Doolani et al. 2020; Partarakis and Zabulis 2024). As highlighted in recent reviews of immersive technologies and AI for human-centered digital experiences (Partarakis and Zabulis 2024), 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 (Alawadh et al. 2024) 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.

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. (2017), 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 (Roldao et al. 2022).

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 (Song et al. 2017; Garbade et al. 2019). Many studies, such as Song et al. (2017), Zhang et al. (2018a, b, 2019), Dourado et al. (2021), Alawadh et al. (2024), use F-TSDF to provide steeper gradients at surface boundaries.

[1] https://unity.com/ (accessed in 2025).

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

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 (Li et al. 2023; Liu et al. 2018; Li et al. 2020, 2019, 2020; Zhong and Zeng 2020). One of the first methods to incorporate RGB features with depth data in the SSC domain was by Liu et al. (2018), where the projection of 2D RGB features is based on depth maps and camera parameters. Other works, such as Cao and de Charette (2022), Wang et al. (2024), Yao et al. (2023), 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 (Garbade et al. 2019; Li et al. 2019; Chen et al. 2020; Cai et al. 2021; Wang et al. 2022; Dourado et al. 2022; Wang et al. 2023). SSC architectures incorporate learning from both 2D and 3D representations leveraged transfer learning to utilise the learnable feature weights from large datasets (Garbade et al. 2019; Chen et al. 2020; Li et al. 2021; Wang et al. 2022; Dourado et al. 2022), with some adopting ResNet-50/ResNet-101 for 2D feature extraction, pre-trained on ImageNet (He et al. 2016; Russakovsky et al. 2015). Research in Garbade et al. (2019), Wang et al. (2022) utilised the pre-trained Deeplab v3+ (Chen et al. 2018) on the ADE20K dataset (2023). A recent study by Wang et al. (2023) employed the Segformer (Xie et al. 2021), initialised with weights from ImageNet. It is noted that some studies have adopted iterative training with distinct learning rates for each input such as Cai et al. (2021), 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 (Wang et al. 2023; Chen et al. 2020; Wang et al. 2022; Tang et al. 2022).

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 Alawadh et al. (2024) 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 Zhang et al. (2019), Li et al. (2023), Liu et al. (2018), 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 (Lee et al. 2024) reconstructs 3D scenes represented by Gaussian splatting from multiple omnidirectional images, while AURORA (Han et al. 2024) 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. (2024) employed CNNs for surface reconstruction from RGB-D, but did not generate annotated 3D models with semantic labels. Earlier methods such as DuLa-Net (Yang et al. 2019) mainly focuses on room layout estimation and struggle with occluded objects, while Meng et al. (2024) proposed a method to recover the room structure without objects semantics from the RGB input. RepFNet (Li et al. 2022) 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. (2019) employed SegNet (Badrinarayanan et al. 2017) 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. (2022) 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 Kim et al. (2022) 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 (Remaggi et al. 2015; Politis et al. 2018; Kim et al. 2022). Several approaches existed for synthesising and generating RIRs (Baran et al. 2024), which can be broadly categorized into algorithmic methods, such as in Raghuvanshi et al. (2010), Lentz et al. (2007), Taylor et al. (2012), and deep learning methods, as in Chen et al. (2023), Liang et al. (2023), Majumder et al. (2022), Ratnarajah et al. (2024), Singh et al. (2021). Some algorithmic methods, like Lentz et al. (2007) and Taylor et al. (2012), 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, Wang et al. (2021) 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 Torres et al. (2004) 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 Shtrepi (2019) showed that the perceptual impact of detailed diffusive surfaces such as triangular prisms on reverberance and spaciousness is noticeably stronger than flat surfaces. Kim et al. (2022) further confirmed that voxelised 3D meshes result in better acoustic realism compared to simpler block-based models, building on earlier findings by Kim et al. (2019). Both studies (Kim et al. 2019, 2022) 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, Kim et al. (2019, 2022) utilised Unity with sound spatialisation plug-ins: Google Resonance Audio[3] in Kim et al. (2019) and Steam Audio in Kim et al. (2022). Notably, Kim et al. (2022) 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 advanced rendering capabilities (Isar 2018; Anil 2024). Unity is particularly noted for its lightweight build and ease of use compared to Unreal Engine (Sabir et al. 2024; Ciekanowska et al. 2021; Isar 2018). Additionally, some studies have explored Unity for spatial sound experiences. For example, Wolf et al. (2020) proposed an optimised binaural sound rendering method for Unity using continuous-azimuth head related transfer functions (HRTFs) to improve localization accuracy. Similarly, Røsvik (2024) 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 (Barron 1995) and RT60 (Rungta et al. 2016). EDT is a metric used to evaluate the acoustics of adjacent reflectors by considering the energy carried by the early reflections (Bradley 2011; Dunn et al. 2015). RT60 is related to the average absorption, location of room boundaries, and room size, describing reverberation from a physical perspective (Bradley 2011; Dunn et al. 2015).

[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).

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 Fig. 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 (Alawadh et al. 2024). 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 Park et al. (2019), Weder et al. (2020). 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 (Song et al. 2017). 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 (Zhang et al. 2019). 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 (Silberman et al. 2012) and NYUCAD (Firman et al. 2016) 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

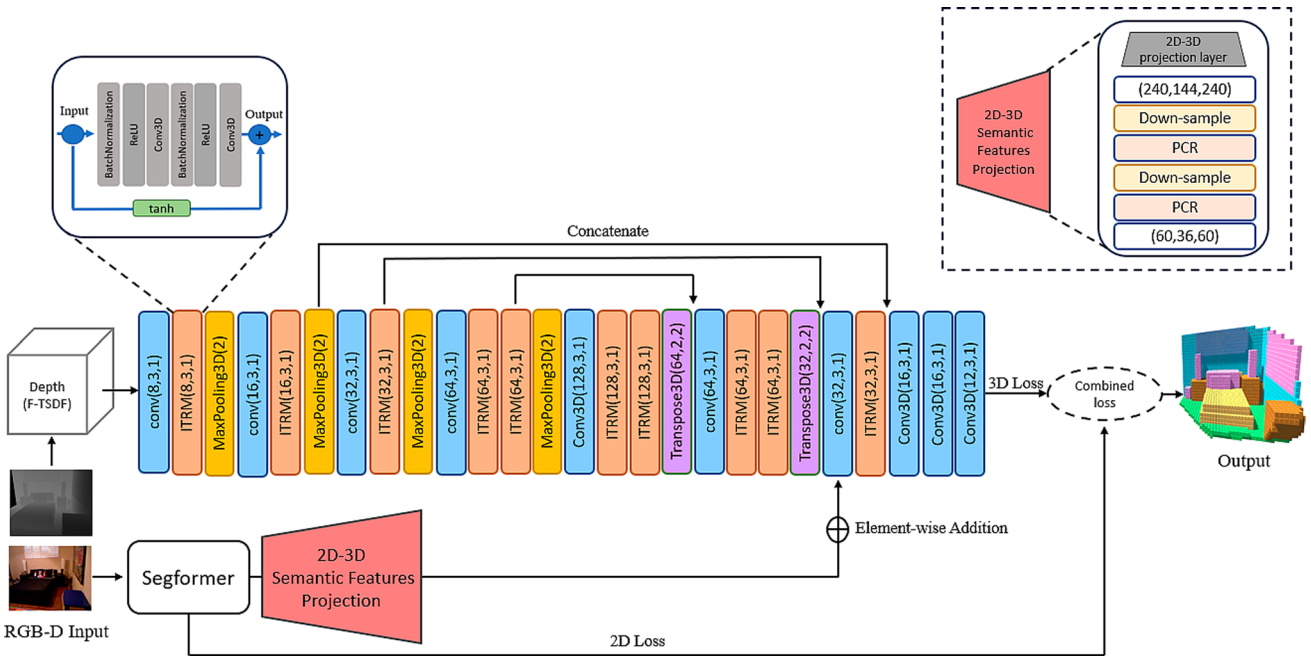


Fig. 2 MDBNet is a dual-head network that processes 2D RGB semantics via a pre-trained Segformer with 2D-3D projection and geometric data via a 3D CNN with ITRM blocks. The network optimises a combined loss, which is a weighted sum of 3D loss and 2D semantics loss

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 (Xie et al. 2021). 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 (Silberman et al. 2012; Firman et al. 2016). 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 (Wang et al. 2023).

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 Liu et al. (2018), 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 (Roldao et al. 2022). 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 (Li et al. 2023), a variant of the Dimensional Decomposition Residual (DDR) block (Li et al. 2019), 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 (Alawadh et al. 2024). The combined loss function is formulated in the following Eq. 1:

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

Aligned with Wang et al. (2023), 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 (Alawadh et al. 2024) 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 Eq. 2:

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

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 Kim and Hilton (2015).

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 Kim et al. (2022). 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) (Pi et al. 2023; Wang 2024; Kim et al. 2022), 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 (Song et al. 2017):

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

The sign in Eq. 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 (Kittler et al. 1998) as illustrated in Fig. 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 Kim et al. (2022). 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 (2007, 2000), Močnik (2023), utilising Eq. 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 highest angular

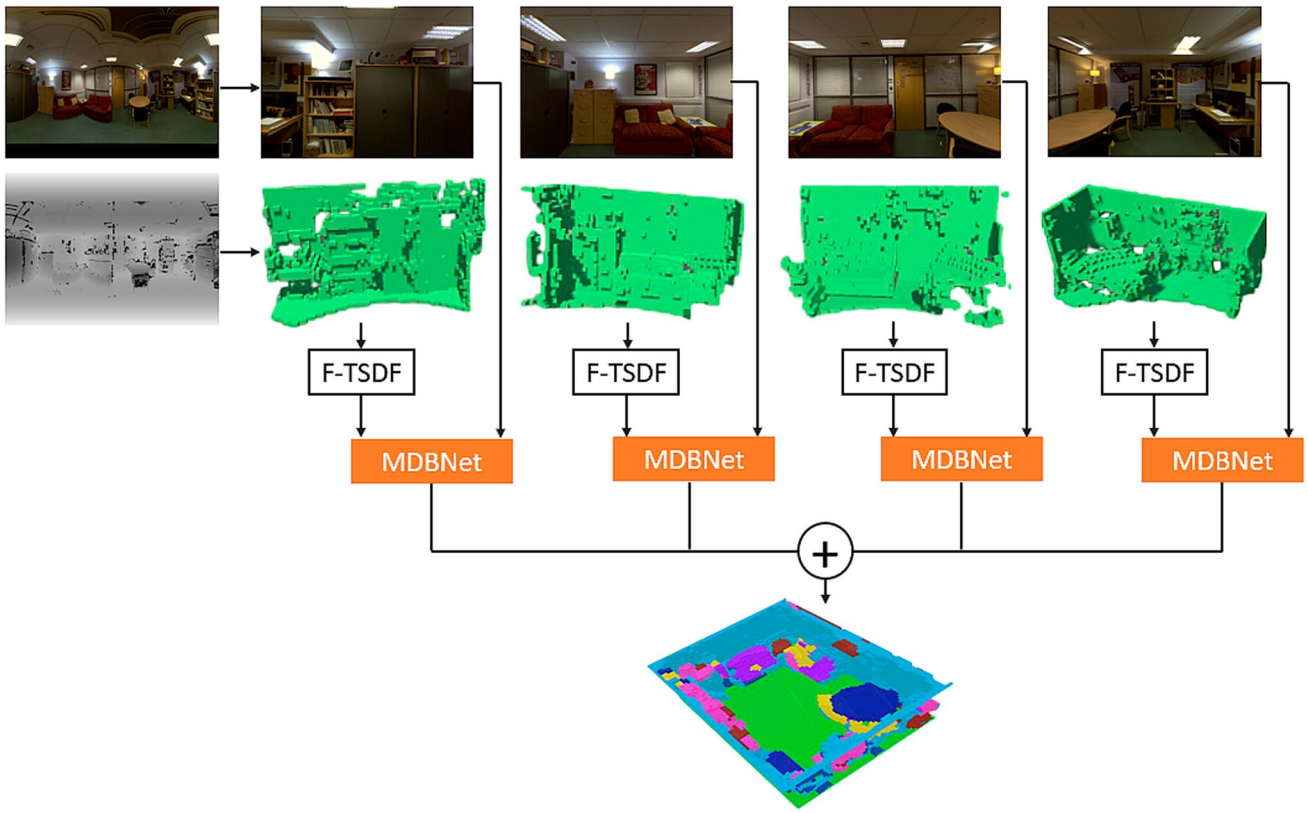


Fig. 3 MDBNet360: RGB-D projection and prediction on full panorama MR scene from CVSSP dataset using MDBNet SSC model

frequency ω_2, as depicted in Eqs. 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 (International Organization for Standardization 2009). This approach employs a linear least-squares fit to determine the slope between 0 dB and -60 dB (Rungta et al. 2016; IoSR 2024). 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 (Barron 1995; IoSR 2024). 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 (Kim et al. 2019, 2022). 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 (Meng et al. 2006) and 5% for EDT (Vorländer 1995).

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 Yao and Mihalcea (2022). For the 2D input modality (RGB), we employ a pre-trained Segformer model (NVIDIA 2024), which is fine-tuned on the ADE20K dataset (2023) 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. (2023), 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 (Alawadh et al. 2024). 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 (Moradi et al. 2020) 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 (Stone 1974; Wong 2015; Rodriguez et al. 2009), 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 Song et al. (2017), 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. (2017). 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 the scene completion area, leading to slight variations among researchers in the field. Some researchers, as seen in Liu et al. (2018) select the occupied occluded voxels while the empty occluded voxels are re-sampled. On the other hand, SPAwN (Dourado et al. 2022) bypasses re-sampling step for empty occluded voxels and evaluates all unoccupied voxels. Other studies, such as PALNet (Li et al. 2019), DDRNet (Li et al. 2019), and AICNet (Li et al. 2020), include all occupied voxels in the scene, combining visible surfaces with occluded regions for scene completion evaluation. In this research, we follow Liu et al. (2018) by evaluating all occluded occupied voxels and re-sampling empty occluded ones. As highlighted in Liu et al. (2024), Li et al. (2020), 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.

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

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 Fig. 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 (Alawadh et al. 2024) by integrating new components and conduct comprehensive experiments to evaluate their contributions, as detailed in Table 2. Initially, we train

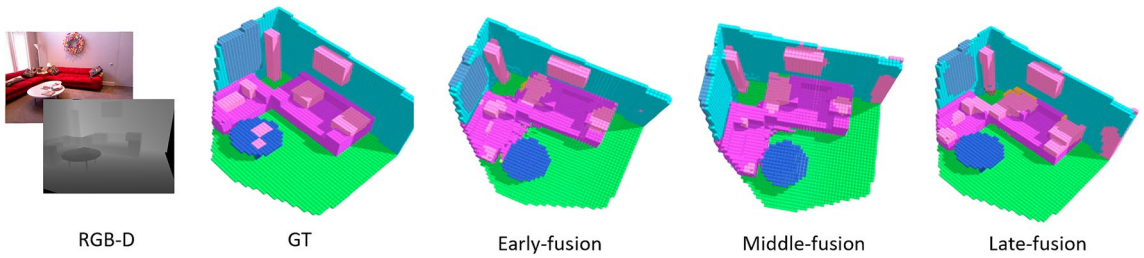


Fig. 4 Illustration of model output using different fusion methods on NYUCAD dataset

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

our model with RGB-D input and apply our combined loss, which includes our proposed re-weighting 3D loss (Alawadh et al. 2024), achieving SSC score of 59.0%. In the second experiment, we replace our proposed re-weighted loss (Alawadh et al. 2024) with a re-sampling-based loss from Song et al. (2017). 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 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 Sect. 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 m. 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 min on a single NVIDIA RTX 8000 GPU. The model utilises a 0.02 m voxel size within a grid of $240 \times 144 \times 240$ for scene input representation, which is scaled down to $60 \times 36 \times 60$ for output, to remain compatible with the MDBNet settings. This representation covers room sizes within $4.8 \times 2.88 \times 4.8$ 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 (Kim et al. 2019, 2022). 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 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 Kim et al. (2022).

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 Kim et al. (2022), this mapping serves as a rough estimation intended to generate plausible sounds and reproducible simulations that allow fair comparison with existing SOTA methods (Kim et al. 2022, 2019). 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

[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

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 Sect. 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 Sect. 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 Sect. 4.1.3, these differences do not necessarily show true performance gaps between SOTA models. Also, Liu et al. (2024), Li et al. (2020) 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 standardised 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, AMMNet$_{Segformer}$ (Wang et al. 2024) which employed Segformer pretrained model for 2D RGB features, and PCANet (Li et al. 2023), respectively. The efficacy of MDBNet is validated on the NYUCAD dataset, as shown in Table 5. Our average performance is competitive with AMMNet$_{Segformer}$ (Wang et al. 2024), and surpasses it when considering the upper-bound results. Furthermore, although our design surpasses SPAwN (Dourado et al. 2022) on the NYUv2 dataset, it demonstrates performance comparable to the more resource-intensive SPAwN model, which utilises semantics priors calculated using surface normals. We observe that scene completion (SC) metrics

Table 4 Results on the NYUv2 dataset include averages and standard deviations for Precision, Recall, IoU, and mIoU metrics

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$_{Segformer}$ Wang et al. (2024)	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 Wang et al. (2023)	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) Cai et al. (2021)	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* Li et al. (2023)	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 Dourado et al. (2022)	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 Alawadh et al. (2024)	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

In the input column, 'D' means depth map only. In the method column, '*' represents the view-volume architecture type

such as Precision and IoU are slightly lower than SOTA methods such as PCANet (Li et al. 2023), AMMNet (Wang et al. 2024), and SISNet (Cai et al. 2021). This is attributed to their explicit architectural emphasis on volumetric occupancy completion, including attention-based modules (PCANet), cross-modal modulation with adversarial training (AMMNet), and scene-instance-scene iteration mechanism (SISNet). MDBNet is primarily oriented toward immersive VR applications, where accurate semantics, particularly surface semantics, play a critical role in material assignment and acoustic rendering. In this context, the completion of deeply occluded volumes has a limited impact on acoustic behaviour.

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 Fig. 5. These comparisons, made between our method and SSCNet (Song et al. 2017), 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 Fig. 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 Fig. 5a, 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 (Song et al. 2017) within the 3D loss, when it applies our proposed re-weighting (Alawadh et al. 2024), and when employing re-weighting (Alawadh et al. 2024) 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 Fig. 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 windows and TVs, showcasing its robustness and adaptability. Additional results are available on our GitHub account: https://github.com/MonaIA1/Repo.

Table 5 Results on the NYUCAD dataset include averages and standard deviations for Precision, Recall, IoU, and mIoU metrics

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$_{Segformer}$ Wang et al. (2024)	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) Cai et al. (2021)	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 Dourado et al. (2022)	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* Li et al. (2023)	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 Alawadh et al. (2024)	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

In the input column, 'D' means depth map only. In the method column, '*' represents the view-volume architecture type

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 (Kim et al. 2022) an extension of EdgeNet (Dourado et al. 2021). 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 Fig. 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 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,

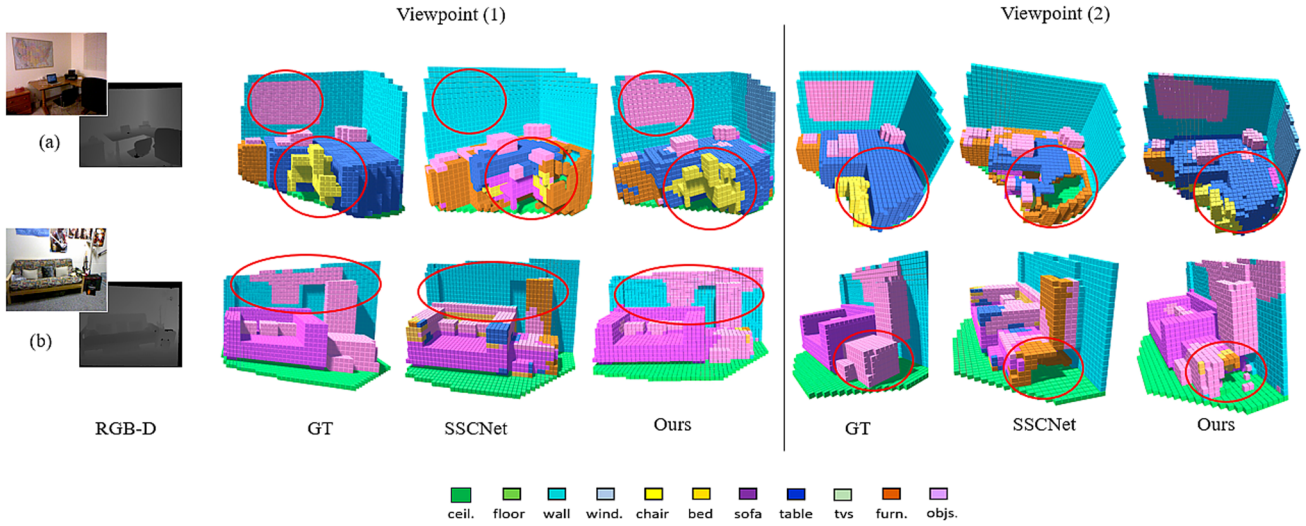


Fig. 5 Comparison of SSC results on the NYUv2 dataset: SSCNet (depth maps) vs. MDBNet (RGB-D). Objects are colour-coded, with circles marking key differences between GT and predictions

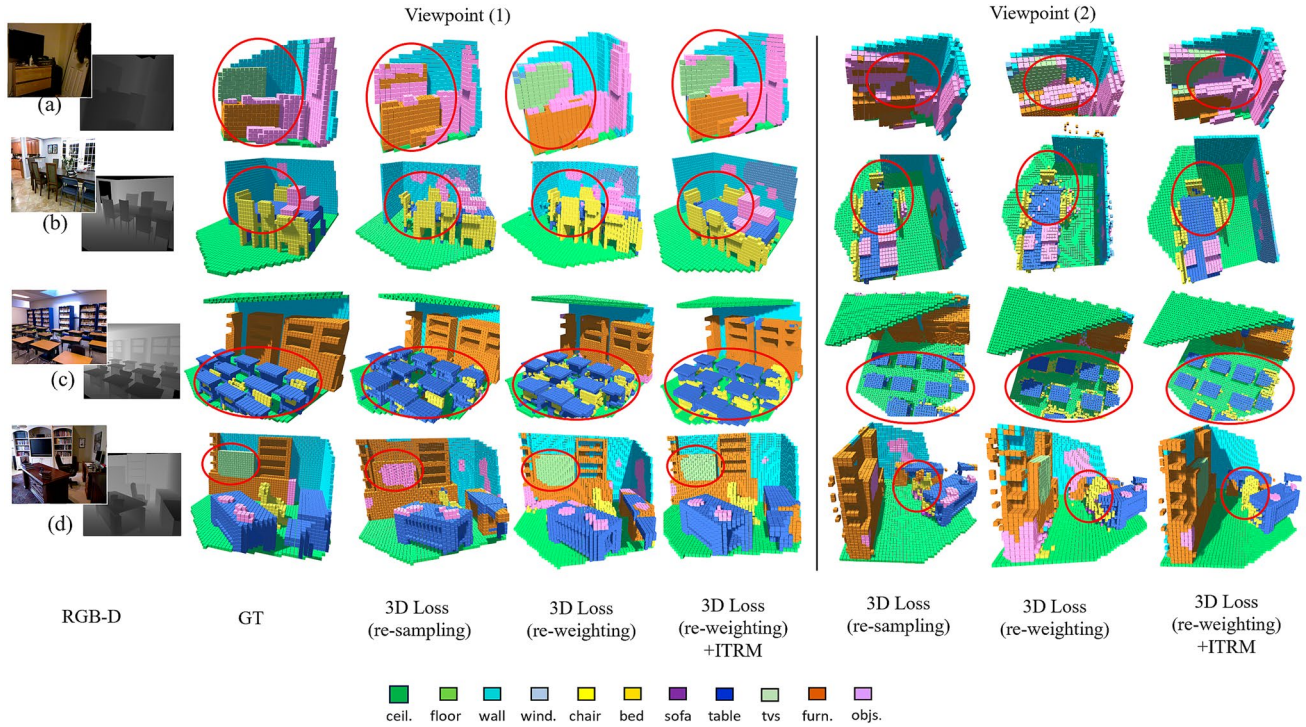


Fig. 6 SSC results with different components on NYUCAD dataset. From left to right: (1) RGB-D input; (2) GT; (3) combined loss with re-sampling; (4) combined loss with re-weighting; (5) combined loss (using re-weighting) with ITRM blocks. Objects are colour-coded, with circles highlighting key differences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

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)

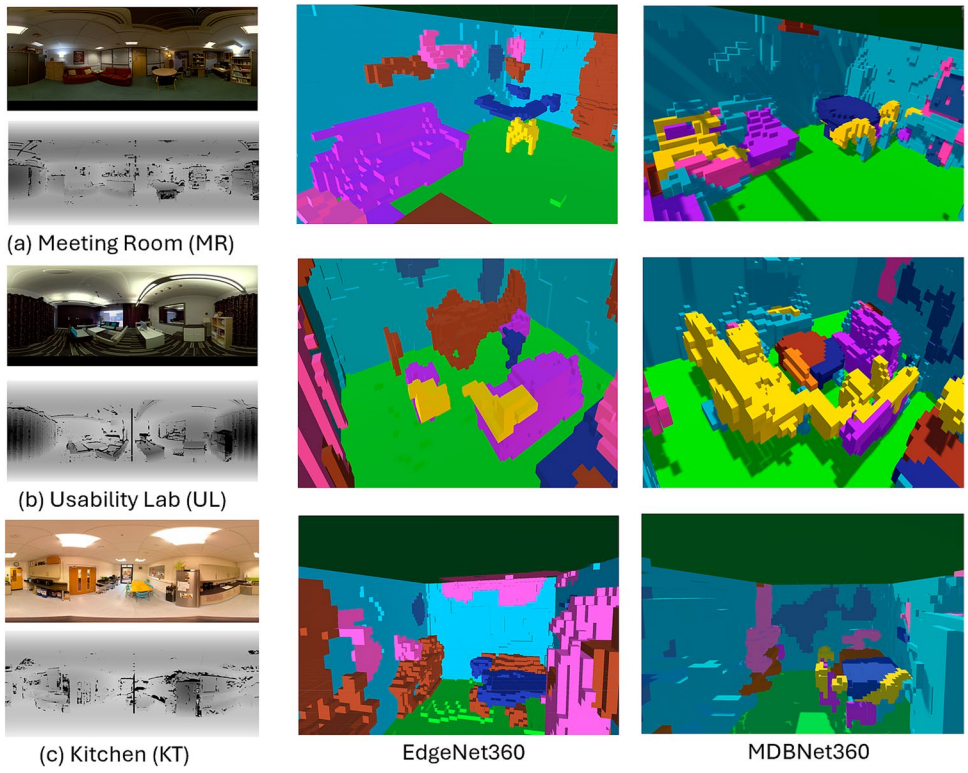


Fig. 7 Qualitative comparison between MDBNet360 and EdgeNet360 on three scenes in CVSSP data. From top to bottom: MR, UL, and KT

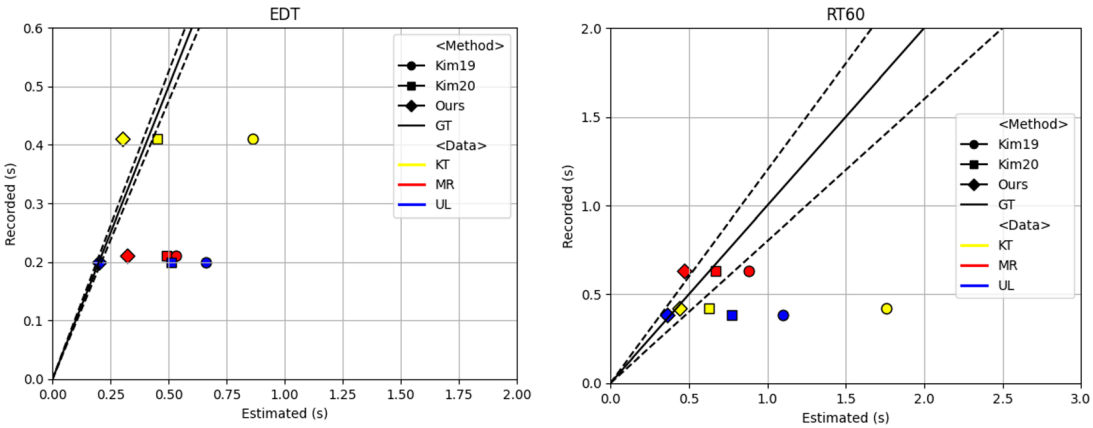


Fig. 8 EDTs and RT60 for 3 CVSSP rooms related to the ground-truth (GT)

measurements obtained from sound modeled in real space, and SOTA models Kim19 (Kim et al. 2019) and Kim20 (Kim et al. 2022). 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 (Meng et al. 2006) and 5% for EDT (Vorländer 1995), 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 Fig. 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 Kim et al. (2020). 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 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 (Alawadh et al. 2024) 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 reweighting 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 (Alawadh et al. 2024), 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 achieved using a late fusion strategy, where learnable features are integrated through downsampling using PCR blocks to match the network's output resolution.

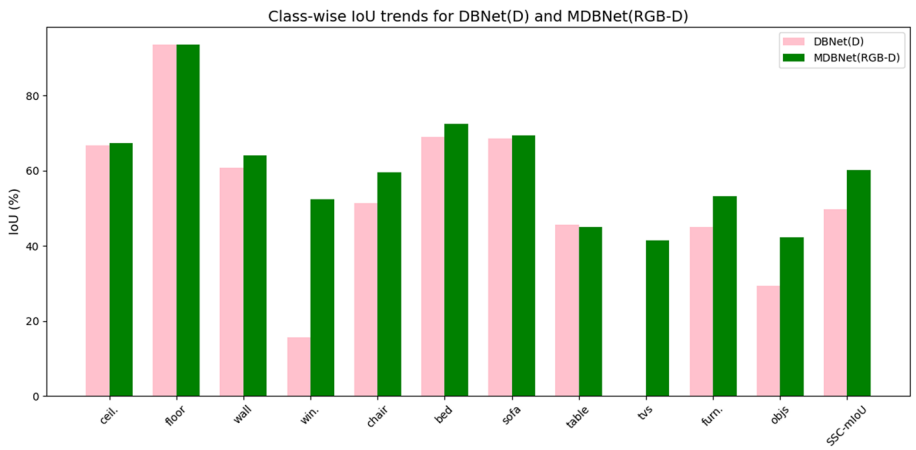


Fig. 9 IoU performance on NYUCAD dataset classes by DBNet model (Alawadh et al. 2024) with depth input and MDBNet model with RGB-D input

This finding is consistent with prior SSC fusion results reported by Roldao et al. (2022). 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 Fig. 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 Sect. 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 Fig. 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 Sect. 5.2, our findings demonstrate that MDBNet360 produced more realistic scene reconstructions and improved semantic completion compared to EdgeNet360 (Kim et al. 2022), 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 (Kim et al. 2019) and Kim20 (Kim et al. 2022). The block-based method in Kim19 (Kim et al. 2019) showed overestimated reverberations due to its simplified, flat surface representations (Shtrepi 2019). Similarly, EdgeNet360's reconstructions Kim20 (Kim et al. 2020) 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 (Wang et al. 2021; Torres et al. 2004).

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.

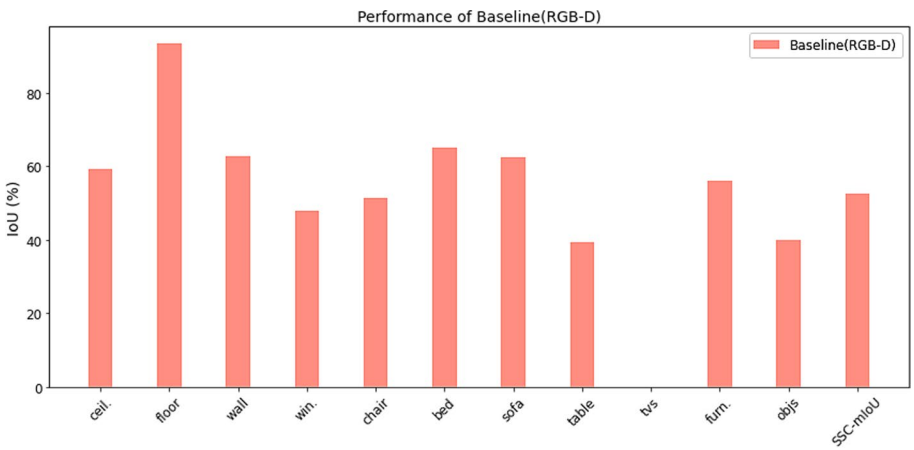


Fig. 10 The baseline architecture performance on semantics level within NYUCAD dataset

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 (Heng et al. 2023) 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 Fig. 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 Fig. 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.

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 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.

[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).

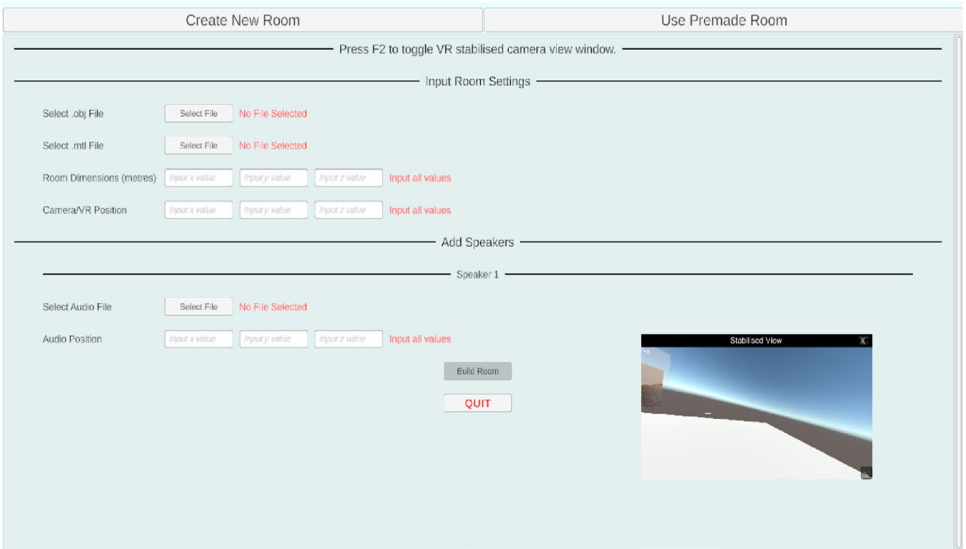


Fig. 12 VR application interface allows users to define and build a custom virtual environment

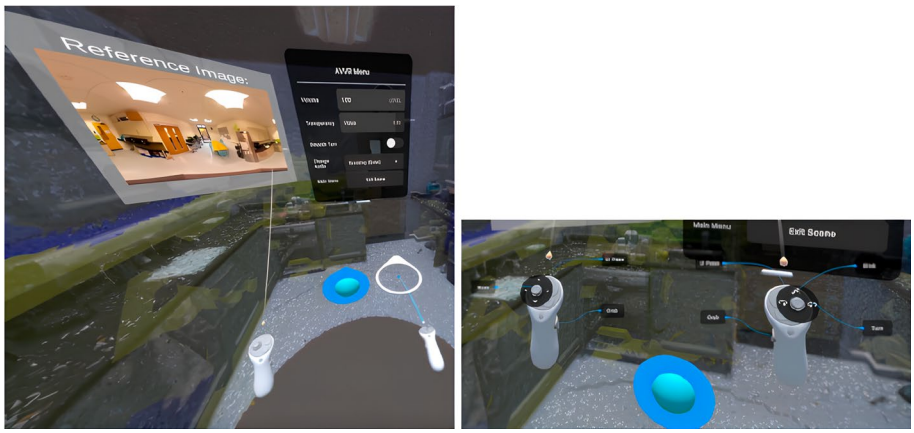


Fig. 13 VR locomotion system showing smooth movement option on the Features menu and the controllers with teleportation

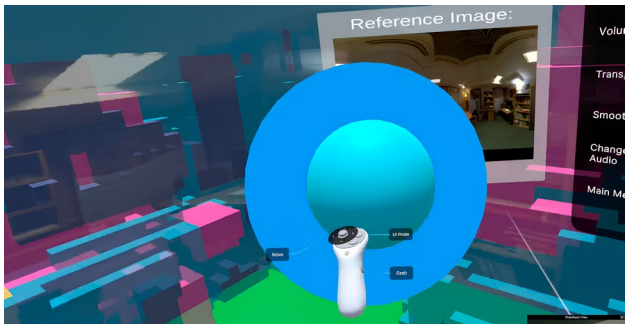


Fig. 14 Illustration of grabbing sound source sphere object (blue) within MR scene

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.

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 Fig. 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 Fig. 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.

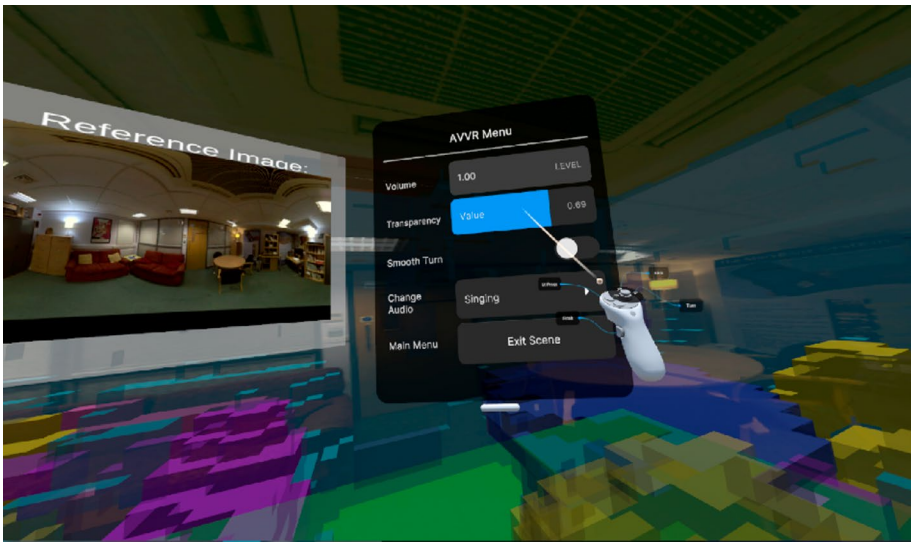


Fig. 15 VR menu showing volume and objects transparency sliders in MR scene

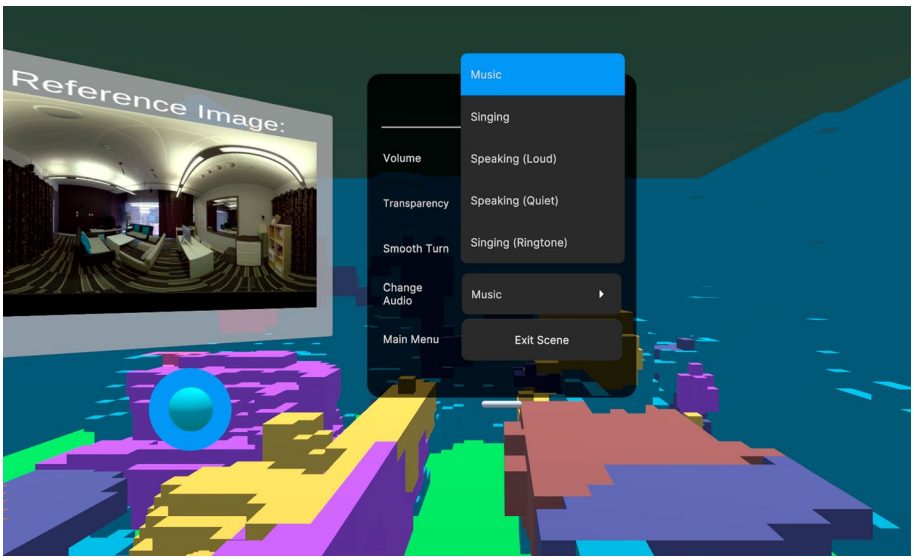


Fig. 16 VR menu showing the audio options in UL scene

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. 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.

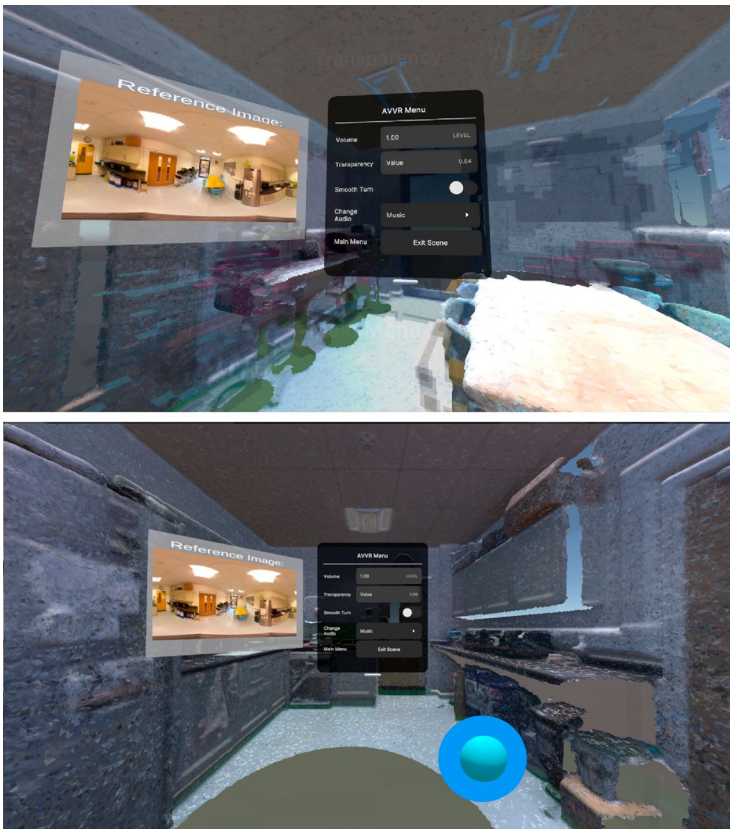


Fig. 17 Illustration of the KT scene with an overlaid LiDAR scan with two different view points

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 questionnaire to identify potential risks such as motion sickness or pre-existing medical conditions that might be aggravated by VR usage (Chang et al. 2020). 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) (Joshi et al. 2015). To complement the quantitative feedback, participants were also encouraged to provide open-ended comments and suggestions, offering qualitative insights into their experiences.

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.

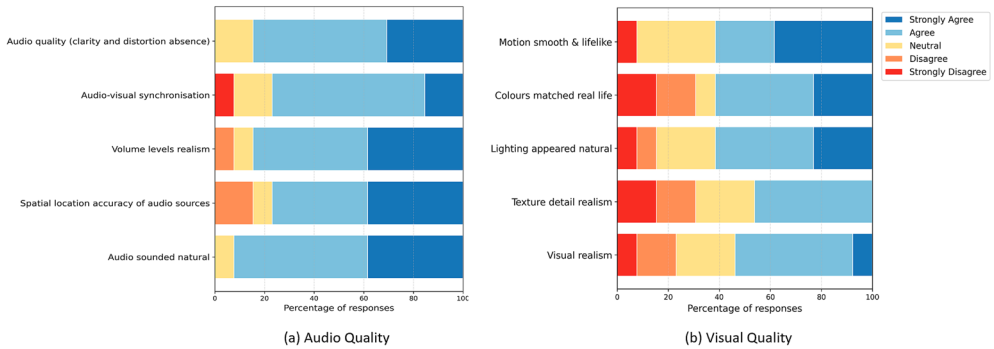
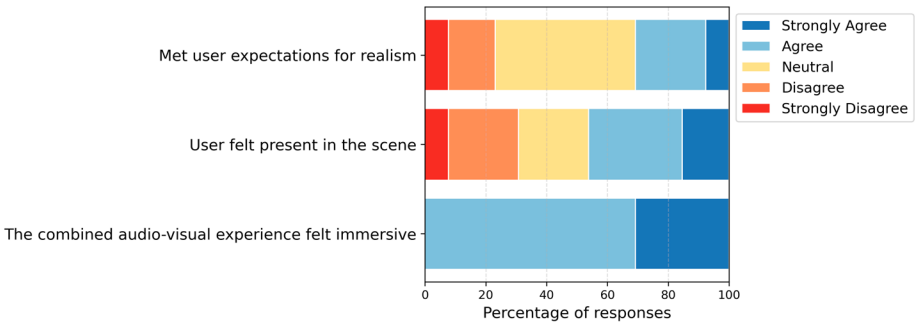


Fig. 18 Responses on evaluating audio and visual aspects in the VR space

Fig. 19 Overall user-experience evaluation



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 Fig. 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 Fig. 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 Fig. 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.

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 (Lin et al. 2017; Ridnik et al. 2021; Cao et al. 2019)) 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 (Chang et al. 2017) and 2D-3D-S (Armeni et al. 2017) 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 Sect. 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.

Author contributions M.A. wrote the main manuscript text. M.A., H.K., and M.N. were involved in the conceptualisation of the study. H.K. and M.N. contributed to the analysis of the model, validation of the results, and interpretation of the findings. A.A. contributed to the evaluation and validation of the spatial sound component.

Funding This work was supported by Electronics and Telecommunications Research Institute (ETRI) grant funded by the Korean government (26ZC1100, Development of Spatial Media Technology and Interaction Technology for Convergence of the Real and Virtual World).

Data availability No datasets were generated or analysed during the current study.

Declarations

Conflict of interest The authors declare no conflict of interest.

Consent to participate 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

ADE20K dataset (2023). https://tinyurl.com/ADE20K. Accessed 17 Jan 2023

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

Anil Ç (2024) 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. Audio Engineering Society

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

Badrinarayanan V, Kendall A, Cipolla R (2017) Segnet: a deep convolutional encoder-decoder architecture for image segmentation. IEEE Trans Pattern Anal Mach Intell 39(12):2481–2495

Baran M-V, King R, Woszczyk W (2024) A general overview of methods for generating room impulse responses. J Acoust Soc Am 155(3–Supplement):282–282

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

Berkman MI (2024) History of virtual reality. In: Lee N (ed) Encyclopedia of computer graphics and games. Springer, Cham, pp 873–881

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

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

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

Cao K, Wei C, Gaidon A, Arechiga N, Ma T (2019) Learning imbalanced datasets with label-distribution-aware margin loss. Adv Neural Inf Process Syst 32

Chang E, Kim HT, Yoo B (2020) Virtual reality sickness: a review of causes and measurements. Int J Human-Comput Interact 36(17):1658–1682

Chang A, Dai A, Funkhouser T, Halber M, Niessner M, Savva M, Song S, Zeng A, Zhang Y (2017) Matterport3d: learning from rgb-d data in indoor environments. arXiv preprint arXiv:1709.06158

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

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

Chen M, Su K, Shlizerman E (2023) 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

Ciekanowska A, Kiszczak-Gliński A, Dziedzic K (2021) Vr space such as. J Comput Sci Inst 20:247–253

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

Dourado A, De Campos TE, Kim H, Hilton A (2021) Edgenet: semantic scene completion from a single rgb-d image. In: ICPR, pp 503–510

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

Dunn F, Hartmann W, Campbell D, Fletcher NH (2015) Springer handbook of acoustics. Springer, New York

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

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

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

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

Han H, Liang Y, Zhou Y, Wang W, J. Rojas-Muñoz E, Li X (2024) 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

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

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

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

IoSR: IoSR Matlab Toolbox. https://github.com/IoSR-Surrey/MatlabToolbox/tree/master. Accessed: 2 Sept 2024

Isar C (2018) A glance into virtual reality development using unity. Inf Economica 22(3):14–22

Joshi A, Kale S, Chandel S, Pal DK (2015) Likert scale: explored and explained. British J Appl Sci Technol 7(4):396

Kim H, Hilton A (2015) Block world reconstruction from spherical stereo image pairs. Comput Vis Image Underst 139:104–121

Kim H, Remaggi L, Jackson PJB, Hilton A (2020) 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. Springer, Cham, pp 293–318

Kim H, Remaggi L, Dourado A, Campos TD, Jackson PJ, Hilton A (2022) Immersive audio-visual scene reproduction using semantic scene reconstruction from 360 cameras. Virtual Real 26(3):823–838

Kim H, Remaggi L, Jackson PJ, Hilton A (2019) 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

Kittler J, Hatef M, Duin RP, Matas J (1998) On combining classifiers. IEEE Trans Pattern Anal Mach Intell 20(3):226–239

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

Lee S, Chung J, Huh J, Lee KM (2024) ODGS: 3d scene reconstruction from omnidirectional images with 3d gaussian splattings. Adv Neural Inf Process Syst 37:57050–57075

Lentz T, Schröder D, Vorländer M, Assenmacher I (2007) Virtual reality system with integrated sound field simulation and reproduction. EURASIP J Adv Signal Process 2007:1–19

Li J, Liu Y, Yuan X, Zhao C, Siegwart R, Reid I, Cadena C (2019) Depth based semantic scene completion with position importance aware loss. IEEE Robot Automat Lett 5(1):219–226

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

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

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

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

Li M, Meng M, Zhou Z (2022) 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

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

Li J, Song Q, Yan X, Chen Y, Huang R (2023) From front to rear: 3d semantic scene completion through planar convolution and attention-based network. IEEE TMM 25:8294–8307

Liu S, Hu Y, Zeng Y, Tang Q, Jin B, Han Y, Li X (2018) See and think: disentangling semantic scene completion 31

Liu X, Xie H, Zhang S, Yao H, Ji R, Nie L, Tao D (2024) 2d semantic-guided semantic scene completion. Int J Comput Vision 133(3):1306–1325

Li T, Zhang Z, Wang Y, Cui Y, Li Y, Zhou D, Yin B, Yang X (2024) Self-supervised indoor scene point cloud completion from a single panorama. Visual Comput 41(3):1891–1905

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

Majumder S, Chen C, Al-Halah Z, Grauman K (2022) Few-shot audio-visual learning of environment acoustics. Adv Neural Inf Process Syst 35:2522–2536

Mandal S (2013) Brief introduction of virtual reality & its challenges. Int J Sci Eng Res 4(4):304–309

Meng M, Zhou Y, Zuo D, Li Z, Zhou Z (2024) Structure recovery from single omnidirectional image with distortion-aware learning. J King Saud Univ-Comput Inf Sci 36(7):102151

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

Močnik M (2023) Pyrirtool: a python tool for room impulse response (RIR) processing. https://github.com/maj4e/pyrirtool. Accessed: 2 Sept 2024

Moradi R, Berangi R, Minaei B (2020) A survey of regularization strategies for deep models. Artif Intell Rev 53(6):3947–3986

NVIDIA: SegFormer B5 Finetuned ADE 640x640. http://tinyurl.com/segformerb5. Accessed: 6 Feb 2024

Pan Y, Xie F, Zhao H (2023) Understanding the challenges when 3d semantic segmentation faces class imbalanced and ood data. IEEE Trans Intell Transp Syst 24(7):6955–6970

Park JJ, Florence PR, Straub J, Newcombe RA, Lovegrove S (2019) Deepsdf: learning continuous signed distance functions for shape representation. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 165–174

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

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

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

Privitera AG, Fontana F, Geronazzo M (2024) The role of audio in immersive storytelling: a systematic review in cultural heritage. Multimedia Tools Appl 84(16):16105–16143

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

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

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

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

Rodriguez JD, Perez A, Lozano JA (2009) Sensitivity analysis of k-fold cross validation in prediction error estimation. IEEE Trans Pattern Anal Mach Intell 32(3):569–575

Roldao L, De Charette R, Verroust-Blondet A (2022) 3d semantic scene completion: a survey. IJCV 130(8):1978–2005

Røsvik PM (2024) Creating a virtual reality orchestral concert experience with 3d audio. Master's thesis, The University of Bergen

Rungta A, Rust S, Morales N, Klatzky R, Lin M, Manocha D (2016) Psychoacoustic characterization of propagation effects in virtual environments. ACM Trans Appl Percept (TAP) 13(4):1–18

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

Sabir A, Hussain R, Pedro A, Soltani M, Lee D, Park C, Pyeon J-H (2024) Synthetic data generation with unity 3d and unreal engine for construction hazard scenarios: a comparative analysis

Shtrepi L (2019) Investigation on the diffusive surface modeling detail in geometrical acoustics based simulations. J Acoust Soc Am 145(3):215–221

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

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

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

Stecker GC, Moore TM, Folkerts M, Zotkin D, Duraiswami R (2018) 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, Audio Engineering Society

Stone M (1974) Cross-validatory choice and assessment of statistical predictions. J Roy Stat Soc: Ser B (Methodol) 36(2):111–133

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

Taylor M, Chandak A, Mo Q, Lauterbach C, Schissler C, Manocha D (2012) Guided multiview ray tracing for fast auralization. IEEE Trans Visual Comput Graph 18(11):1797–1810

Torres R, Rycker N, Kleiner M (2004) Edge diffraction and surface scattering in concert halls: physical and perceptual aspects. J Temp Design Architect Environ 4(1):52–58

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

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

Wang Y (2024) Projection methods for 360-degree video. Front Comput Intell Syst

Wang X, Feng W, Wan L (2024) Multi-modal fusion architecture search for camera-based semantic scene completion. Expert Syst Appl 243:122885

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

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

Wang R, Zhang Y, Jia B (2021) 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), IEEE, pp 345–349.

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

Weder S, Schönberger JL, Pollefeys M, Oswald MR (2020) Neuralfusion: online depth fusion in latent space. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 3161–3171

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

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

Xie E, Wang W, Yu Z, Anandkumar A, Alvarez JM, Luo P (2021) Segformer: simple and efficient design for semantic segmentation with transformers. Adv Neural Inf Process Syst 34:12077–12090

Yang S-T, Wang F-E, Peng C-H, Wonka P, Sun M, Chu H-K (2019) 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

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

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

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

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

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

Zhong M, Zeng G (2020) 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. IOS Press, Amsterdam, pp 2824–2831