UNIVERSITY OF SOUTHAMPTON

Faculty of Engineering and Physical Sciences
School of Electronics and Computer Science

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

DOI: 10.1002/0470841559.ch1

Volume n of m

by

Mona Ibrahim Alawadh

Supervisors:
Dr. Hansung Kim
Prof. Mahesan Niranjan

ORCiD: 0000-0002-1825-0097

A thesis for the degree of
Doctor of Philosophy

June 2025

University of Southampton

Abstract

Faculty of Engineering and Physical Sciences
School of Electronics and Computer Science

Doctor of Philosophy

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

by Mona Ibrahim Alawadh

In this research, we propose a novel method for generating an audio-visual scene in 3D virtual space using a single panoramic RGB-D input. Our investigation begins with the reconstruction of a 3D model from RGB panoramic data alone, developing a semantic geometry model by combining estimated monocular depth with material information for spatial sound rendering. Building upon the preliminary results, we extend our approach to construct a comprehensive virtual reality (VR) environment using 360° RGB-D input. The proposed method enables the creation of an immersive VR space by generating a complete 3D voxelized model that incorporates scene semantics from a single panoramic input.

Our methodology employs a deep 3D convolutional neural network integrated with transfer learning for RGB semantic features, coupled with a re-weighting strategy in the 3D weighted cross-entropy loss function. The proposed re-weighting method uniquely combines two class re-balancing techniques (re-sampling and class-sensitive learning) while smoothing the weights through an unsupervised clustering algorithm. This approach addresses critical challenges in semantic scene completion (SSC), including inherent class imbalances in indoor 3D spatial representations. Furthermore, we quantify the performance uncertainty in our results to ensure an unbiased assessment across trials, contributing to more reliable benchmarking in the SSC field. We design a hybrid architecture featuring a dual-head model 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. RGB features are projected from 2D to 3D space using depth maps. We explored various RGB semantics fusion strategies, including early, middle, and late fusion methods. Based on performance evaluations using K-fold cross-validation, we selected the late fusion approach. This method involves downsampling features using planar convolutions to align with 3D resolution, followed by fusing RGB semantic features with geometric information through element-wise addition. The hybrid encoder-decoder architecture incorporates

an Identity Transformation within a full pre-activation Residual Module (ITRM), enabling effective management of diverse signals within the F-TSDF representation.

The inference methodology of the proposed SSC model is extended to accommodate 360° RGB-D input through cubic projection and 3D rotation, enabling VR space design with comprehensive spatial coverage. We propose a streamlined computer vision-based approach capable of reconstructing a 3D SSC model from a single panoramic input, facilitating plausible sound environment simulation. Additionally, our proposed method contributes to reducing the complexity of estimating room impulse responses (RIRs), which typically require extensive equipment and multiple recordings in real space. We implement the audio-visual VR reconstructions in the Unity 3D gaming platform combined with the Steam audio plug-in for spatial sound rendering. Acoustic properties are evaluated by measuring parameters such as early decay time (EDT) and reverberation time (RT60). Comparative analysis indicates that our approach achieves better VR space reconstruction, producing more realistic scene representations and immersive acoustic characteristics compared to existing methods reported in the literature.

The proposed method contributes to the design of enhanced VR environments by integrating both audio and visual signals into a unified framework. Our results support the development of datasets that combine audio and 3D SSC models, encouraging the application of AI in VR spaces. This advancement has the potential to drive progress in VR applications across various domains, such as gaming, education, and tourism.

Contents

List of Figures

List of Tables

Declaration of Authorship

I declare that this thesis and the work presented in it is my own and has been generated by me as the result of my own original research.

I confirm that:

1. This work was done wholly or mainly while in candidature for a research degree at this University;
2. Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
3. Where I have consulted the published work of others, this is always clearly attributed;
4. Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
5. I have acknowledged all main sources of help;
6. Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
7. Parts of this work have been published as:
    - Room Acoustic Properties Estimation from a Single 360◦ Photo. Published in European Signal Processing Conference (EUSIPCO) 2022.
    - 3D Semantic Scene Completion From a Depth Map with Unsupervised Learning For Semantics Prioritisation. Published in International Conference on Image Processing (ICIP) 2024.
    - MDBNet 360°: 3D Audio-Visual Indoor Scene Reconstruction and Completion from a Single 360° RGB-D Image. Accepted in 3D Workshop on Computer Vision for Mixed Reality (CV4MR) 2025.

Signed: Date:

Acknowledgements

All praise and gratitude are due to Allah, whose grace and blessings have enabled me to complete this dissertation. This journey would not have been possible without His guidance and support.

I extend my heartfelt thanks to my dear parents, whose encouragement and prayers have been a constant source of strength. Their unwavering faith in me has been a profound inspiration, even across distances.

To my beloved husband, Yazeed, I express my deepest gratitude for your unwavering support. Your belief in me and encouragement have been invaluable, and I could not have accomplished this without you by my side.

To my wonderful children, Badir, Reem, Saud, and Aljazi, thank you from the depths of my heart. Your love, understanding, and encouragement have been my greatest motivation. Every moment spent with you renewed my strength and determination to persevere. I strive to make you proud in every step I take.

To my siblings and extended family, your words of encouragement and constant presence have been invaluable. I am deeply grateful for the love and support you have shown me.

I am also immensely grateful to my friends, particularly those who supported me during my time abroad. My special thanks go to Bashaier Al-Nassir, Reem Al-Hajji, Sarah Abdulmalik, Khulood Al-Ghamdi, Sarah Al-Ahmadi, Lulwa Al-Sanea, Shatha Al-Qarni, Sarah Al-Sudairi, and Maha Al-Thiyabi for your unwavering kindness and support.

To my research collaborators and friends, Daniela Mihai, Atiyeh Alinaghi, Yihong Wu, and Yuwen Heng, thank you for your support and contributions throughout this journey.

Finally, I wish to express my profound appreciation to my supervisors, Dr. Hansung Kim and Prof. Mahesan Niranjan. Your belief in my abilities, guidance, and constructive feedback have been fundamental to the development of my work and growth as a researcher. I am truly honoured to have had your mentorship.

This dissertation is not merely words on paper but the result of dedicated effort, continuous encouragement, and the love of those around me. All my thanks and gratitude to everyone who played a role in this journey.

Chapter 1

Introduction

"Any appearances whatever present themselves, not only when its object stimulates a sense, but also when the sense by itself alone is stimulated, provided only it be stimulated in the same manner as it is by the object."

Aristotle, 330 B.C.

The process of perceiving information from available stimuli is a fundamental aspect of the human perceptual system (Gibson, 1966). Both auditory and visual systems play critical roles in interpreting our surroundings, enabling us to navigate and understand complex environments. In virtual reality (VR), 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). Many domains use VR applications such as education, gaming, tourism, and engineering (Kavanagh et al., 2017; Mandal, 2013; Berni and Borgianni, 2020; Guttentag, 2010; Bretos et al., 2024). Immersion refers to the creation of powerful illusions of reality that rely on the degree to which high-fidelity inputs, such as light patterns and sound waves, are delivered to sensory modalities like vision, audition, and touch (Mandal, 2013; Berkman, 2024). Among these, the immersion effect is primarily based on visual perception (Berkman, 2024).

Building on the role of visual perception in immersive experiences, this thesis 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 3D models with semantics for use in VR spaces. Since 2D images capture only partial information about 3D scenes, AI enables the development of models capable of understanding and learning the underlying structure and semantics of the 3D world, including the reconstruction of occluded areas from a single 2D input. Specifically, this research focuses on Semantic Scene Completion (SSC), a challenging

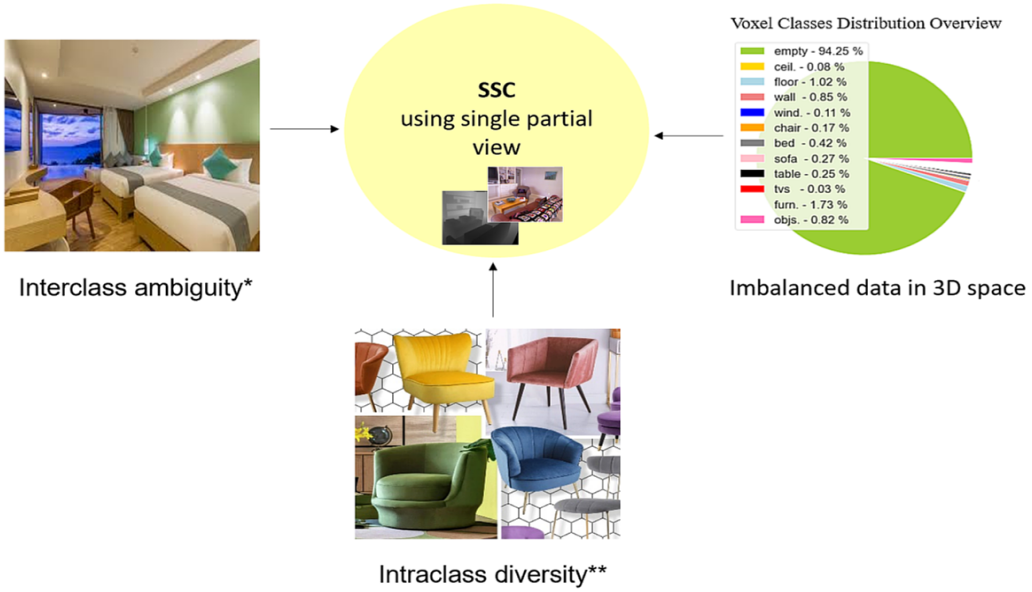

FIGURE 1.1: SSC Challenges. '*' Image source: crestvillas [1], and '**' Image source: stylistchair [2].

task in computer vision that involves predicting 3D annotated models, including occluded regions, from single-perspective views of indoor scenes. SSC is an ill-posed problem for voxelized indoor scenes, with the primary challenge being the prediction of a complete 3D representation from the inherently limited information available in a single 2D image. Due to the partial-view nature of the input, SSC involves significant loss of 3D information in occluded regions. 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). Intraclass diversity arises when objects within the same semantic category (y) exhibit significant variability in sensor data (x); for example, chairs may vary widely in shape and size, creating a substantial semantic gap. Interclass ambiguity, on the other hand, occurs when objects from different semantic categories (y) appear similar due to partial observations or occlusions; for instance, a sofa might be misinterpreted as a bed, highlighting the sensory gap. Figure 1.1 illustrates these primary challenges in SSC. Furthermore, 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 input. Constructing 3D models from partial views alone often falls short of providing the fully immersive experience required for realistic VR applications. To address this limitation, this study aims to

[1] https://www.hotelspatongthailand.com/en/property/crest-resort-pool-villas.html (accessed in 2024)

[2] https://www.stylist.co.uk/home/affordable-velvet-chair/564984 (accessed in 2024)

generate complete VR spaces with 360° surroundings, creating an environment that mirrors the user's spatial perception in the real world.

To achieve enhanced immersion, spatial sound must be integrated with 3D models. In this research, sound rendering and modeling are performed using a VR gaming engine equipped with spatial sound plug-ins to generate a 3D digital 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 parameters such as early decay time (EDT) and reverberation time (RT60) using the exponential sine sweep method (ESS).

In addition to advancing machine understanding through the reconstruction of 3D environments from 2D inputs and the integration of scene acoustics to create immersive experiences, this research contributes to bridge the gap between computational modeling and human perception. It introduces a horizontal integration of AI and VR, designed to support more intuitive, human-centered digital interactions. By merging spatial sounds and spatial scene semantics in the 3D space, the system enables users to experience virtual environments that feel both intelligible and engaging. 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). The VR application proposed in this research demonstrates how users can actively interact with and engage in reconstructed 3D spaces, thereby fostering a deeper spatial understanding. 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.

1.1 Research Questions

This research addresses the following research questions (RQs):

1. RQ 1: How can we generate an audio-visual VR space from a single 360 ° RGB input?
2. RQ 2: How can the inference capabilities of pre-trained SSC model on perspective images be extended to a single 360° RGB-D input?

   To answer the second research question RQ 2, sub-questions exist and are related to the design of SSC models. These questions should be addressed first to answer RQ 2.

   (a) RQ 2a: What is the impact of prioritising voxel weights within the scene on the SSC model learning?
   (b) RQ 2b: What is the impact of learning multiple features from RGB-D input on the performance of the proposed SSC model?

3. RQ 3: What is the impact of the 3D scene generated from the 360° RGB-D input on the acoustic parameters, including early reflections and late reverberations?

The following section outlines the main contributions of this research, linking these contributions to the research questions and the corresponding chapters of this thesis.

1.2 Contributions

Through the systematic design and integration of multiple methodological components, this thesis makes several contributions that advance the current understanding of the research domain. In the following sections, we present the main contributions and their related sub-contributions.

1. **Constructing a full VR space with acoustic materials from a single 360° RGB input**. This contribution addresses the first research question RQ 1 by proposing a pipeline that integrates multiple outputs, including mono depth and material estimation from a single 360° RGB input. These components are then processed to construct a 3D SSC model with integrated sound rendering. Preliminary results are presented in Chapter 3, which also provide the motivation and foundational roadmap for the subsequent chapters in this research.

   - Propose a cubic projection on material estimation input to decompose the 360° RGB into perspective inputs compatible with the material estimation network.
   - Evaluate the room acoustic parameters in a virtual space, such as EDT and RT60.

   These contributions were published in the 30th European Signal Processing Conference, (EUSIPCO) 2022.

2. **Designing a re-weighting method in the cross entropy loss based on unsupervised learning using clustering algorithms**. Investigate different re-weighting strategies and design re-weighting method for cross entropy loss uniquely combines two class re-balancing approaches (re-sampling and class-sensitive learning) and smooths the weights using unsupervised clustering algorithms. This contribution addresses the data imbalanced challenge as depicted in Figure 1.1, and answered the research question RQ 2a. Details are provided in Chapter 4.

- Introduce the performance uncertainty by calculating average scores along with standard deviations using K-fold cross-validation. This method ensures an unbiased assessment across trials and contributes to more reliable benchmarking in the SSC field.
- Achieve state-of-the-art (SOTA) performance in scene semantics completion task, significantly surpassing other comparable methods on two public benchmark datasets using only single-depth input.

These contributions were published in the 2024 IEEE International Conference on Image Processing, (ICIP) 2024.

3. **Designing MDBNet model an enhanced 3D SSC model from single perspective RGB-D input with an employment of transfer learning**. A dual-head architecture was developed to simultaneously train the model using combined loss function for both perspective RGB and 3D SSC. This contribution addresses research question RQ 2b and reduces class ambiguity, as illustrated in Figure 1.1. Further details are provided in Chapter 5.

    - Evaluate different RGB semantics fusion strategies by assessing the average performance scores using K-fold cross-validation. This comprehensive analysis facilitates the selection of fusion method that effectively validate the model's generalisation across diverse scenarios.
    - Enhance the overall results by implementing the Identity Transformed within the full pre-activation Residual Module (ITRM), with a hyperbolic tangent activation function applied to identity features. Also, we analyse the impact of learning multiple features on the proposed SSC model.

4. **Designing MDBNet360 by Extending MDBNet's inference capabilities and providing a comprehensive 3D model prediction from single 360° RGB-D input**. Building upon the SSC model MDBNet developed in Chapter 5, this research extended its predictive capabilities by leveraging cubic projection and 3D rotation techniques. The methodology enables the reconstruction of a comprehensive 3D model with full spatial surroundings from a single 360° RGB-D input, originally pre-trained on perspective RGB-D datasets. This methodological extension directly addresses research question RQ 2, with comprehensive details elaborated in Chapter 6.

    - Conduct a qualitative assessment to evaluate the reconstructed scenes and compare the results with SOTA approaches.

5. **Constructing an immersive VR Space by integrating audio-visual data using MDBNet360**. We rendered an exponential sine sweep (ESS) within the reconstructed scenes by MDBNet360 using the Unity game engine and Steam Audio plugin for advanced spatial sound rendering. The comprehensive analysis of the

acoustic parameters reveals the relationships between acoustic parameters and the predicted 3D models, directly addresses research question RQ 3. Detailed methodological insights are elaborated in Chapter 6.

- Perform an acoustic analysis of the 3D virtual environments generated by MDBNet360 through detailed evaluation of RIRs acoustic parameters, specifically examining EDT and RT60, and comparing the results with SOTA methods.
- Develop a VR software application in collaboration with graduate students in the School of Electronics and Computer Science (ECS) at the University of Southampton to demonstrate the proposed method in a VR setting.

1.3 Publications

The contributions of this PhD research, published or under review at peer-reviewed conferences, are listed below:

1. Room Acoustic Properties Estimation from a Single 360° Photo. *Published in European Signal Processing Conference (EUSIPCO) 2022.*
2. 3D Semantic Scene Completion From a Depth Map with Unsupervised Learning For Semantics Prioritisation. *Published in International Conference on Image Processing (ICIP) 2024.*
3. MDBNet 360°: 3D Audio-Visual Indoor Scene Reconstruction and Completion from a Single 360° RGB-D Image. *Accepted in 3D Workshop on Computer Vision for Mixed Reality (CV4MR) 2025.*

1.4 Thesis Structure and Components

This thesis is organized into seven chapters. Chapter 1 introduces the research theme, including the challenges, research questions, and contributions. Chapter 2 establishes the conceptual and theoretical foundations along with related works relevant to this doctoral research. Chapter 3 proposes a pipeline for generating VR space from a single panoramic RGB input. In Chapter 4, a novel loss function design is introduced, based on unsupervised clustering within a SSC model to produce a comprehensive 3D model with scene semantics. Consequently, the proposed model in Chapter 4 is enhanced by integrating additional features from 2D RGB inputs, using combined loss for model training, and providing a comprehensive qualitative and quantitative evaluation presented in Chapter 5. Chapter 6 proposes a novel method for constructing a digital 3D

space within a VR environment using a single 360° RGB-D input. Chapter 7 presents the research findings, and proposes potential research directions for future scholarly investigation. The research workflow is comprehensively depicted in Figure 1.2.

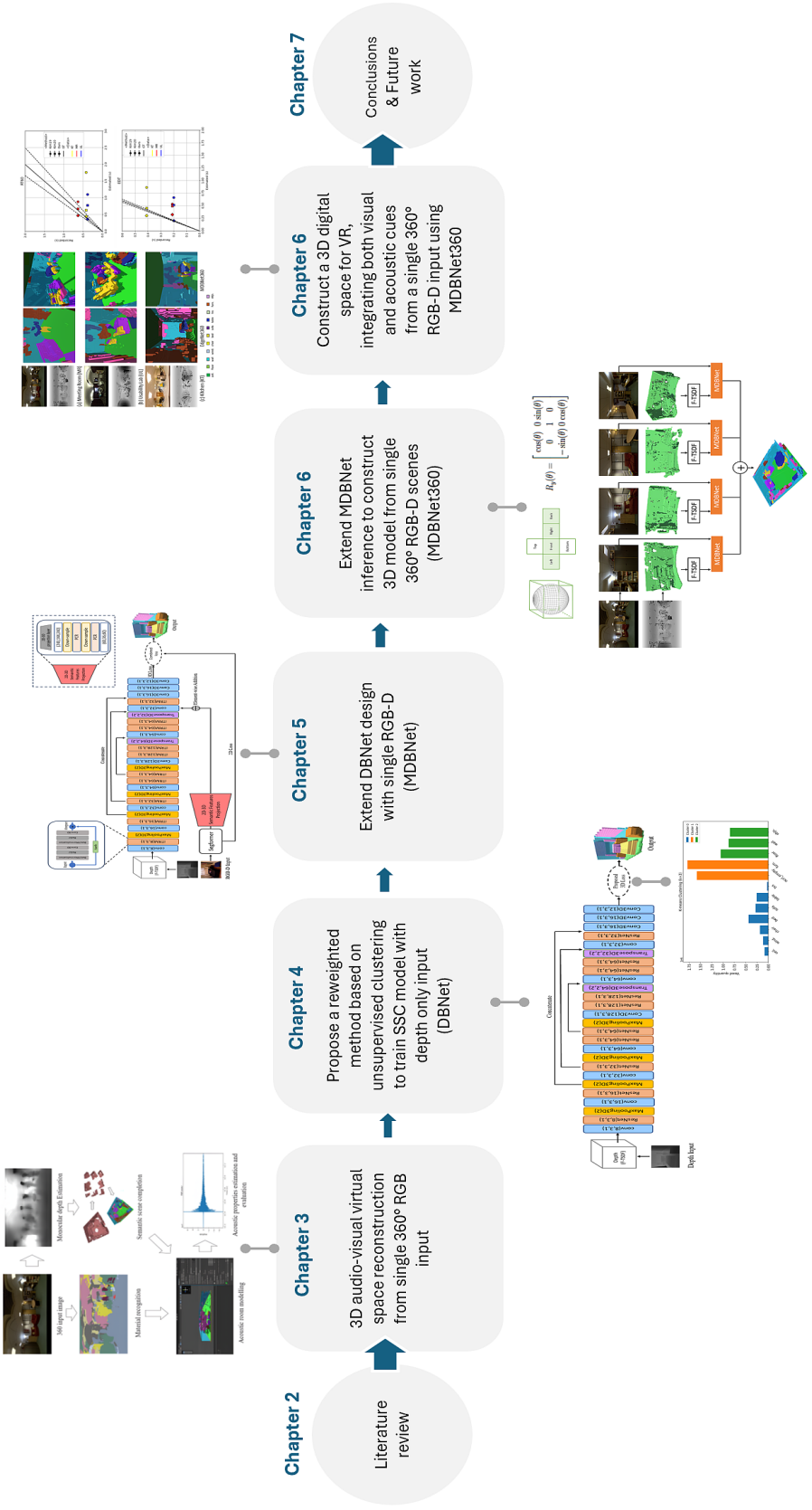

FIGURE 1.2: PhD thesis workflow.

Chapter 2

Background and Literature Review

"Such objects are always imagined as being present in the field of vision as would have to be there in order to produce the same effect on the nervous mechanism."

Hermann Von Helmholtz, 1867.

To design an immersive virtual experience, it is essential to understand the human auditory and visual systems, 3D models, and spatial sounds, and their interaction with each other. This chapter begins with an exploration of the human auditory system and the digital representation of audio in 3D space. Section 2.2 elaborates on human visual system and the representation of visual data in 3D spaces. Section 2.3 discusses the application of deep learning for learning 3D data and semantic scene completion in 3D spaces. Finally, Section 2.4 and Section 2.5 review methods for constructing comprehensive 3D models integrated with audio in VR environments, focusing on inputs derived from single 360° images.

2.1 Spatial Audio

This section explores the foundational concepts and methods that are most related to the audio component of our research.

2.1.1 How Do Humans Hear the Sounds?

While the ear is commonly recognized as the organ for hearing, sound perception is more than just hearing; it involves active listening, which depends on both ears and the muscles that help direct them toward sound sources (Gibson, 1966). The perception of sound is a complex process that is shaped by the physiology of the auditory

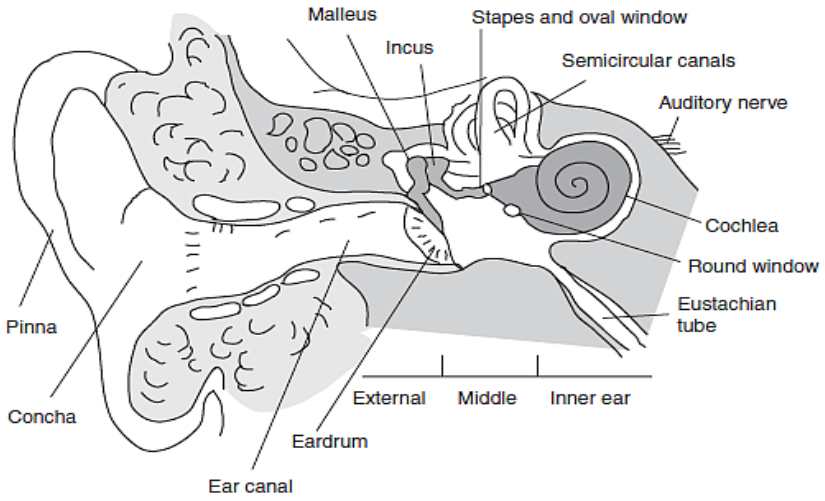

FIGURE 2.1: Diagram of one ear, showing the external, middle, and inner ear. Image sourced from: (Pulkki and Karjalainen, 2015).

system and influenced by cognitive processes. The auditory system processes essential sound properties, such as spectral content, temporal features, and spatial positioning into specific neural patterns that shape our experience of pitch, volume, timbre, and sound location (Roginska and Geluso, 2018). The auditory system, which runs from the ears to the brain's frontal lobes, manages increasingly complex processes as signals ascend through the nervous system. Its roles are typically categorized into peripheral and central processing functions (Roginska and Geluso, 2018). The peripheral auditory system includes the external ear for capturing sound, the middle ear for transmitting vibrations, and the inner ear for converting these vibrations into neural signals. These signals travel through the auditory pathway to the auditory cortex, where higher-level sound processing occurs using the central processing functions (Pulkki and Karjalainen, 2015).

2.1.1.1 Ear Structure

External Ear. The external ear consists of three main components: the pinna with concha, the ear canal (meatus), and the eardrum (tympanic membrane) (Gibson, 1966; Pulkki and Karjalainen, 2015; Roginska and Geluso, 2018), as shown in Figure 2.1. The pinna plays a crucial role in directing high-frequency sound by creating asymmetry in the front-back and top-down directions, enhancing directional hearing. The concha, which connects to the ear canal, collaborates with the head and outer ear, extending to the eardrum, amplifies the sound pressure by 30 to 100 times for frequencies around 3 kHz. This enhancement occurs due to the passive resonance effect resulting from the ear canal's length. As a result, humans are highly sensitive to frequencies between 2–5 kHz, particularly for sound sources directly in front of them in open environments. Additionally, the perception of sound based on its angle of arrival is influenced by

the Head-Related Transfer Function (HRTF), which explains how sound waves interact with the ear and head's shape (Pulkki and Karjalainen, 2015; Roginska and Geluso, 2018).

Middle Ear. The middle ear, located between the eardrum and the inner ear as depicted in Figure 2.1, acts as a mechanical system that transmits vibrations from the eardrum to the cochlea through three tiny bones known as the ossicles: the hammer, anvil, and stirrup. Its complex structure helps match the impedance between air and the fluid in the cochlea, preventing sound energy loss. The difference in size between the eardrum and the oval window further amplifies sound pressure, improving efficiency by around 30 dB. Additionally, the middle ear has an acoustic reflex that protects the inner ear from loud sounds by reducing sound transmission via the ossicles, though this reflex primarily affects low frequencies (less than 1 kHz) reaching the cochlea (Pulkki and Karjalainen, 2015; Roginska and Geluso, 2018; Rossing, 2007).

Inner Ear. The main component of the inner ear responsible for hearing is the cochlea as shown in Figure 2.1, which transforms mechanical sound vibrations from the middle ear into neural signals through a process known as mechanoelectrical transduction. This occurs in the basilar membrane, a key structure within the cochlea, which is a spiral, liquid-filled tube. The organ of Corti, located on the basilar membrane, contains sensory hair cells that are essential for converting mechanical stimuli into electrochemical signals. Acoustical waves create a traveling wave along the basilar membrane, causing the hair bundles (stereocilia) on the hair cells to bend. This bending generates electrical signals, allowing the auditory nerve fibers to encode the frequency, amplitude, and phase of the sound. Due to variations in the stiffness of the basilar membrane—where the base is stiffer than the apex—high-frequency sounds (20 kHz) cause displacement near the base, while low-frequency sounds (20 Hz) displace the membrane near the apex. This arrangement allows the basilar membrane to function as a series of overlapping filters, with each region corresponding to a specific characteristic frequency. Hair cells in each region are "tuned" to these frequencies, resulting in a spatial organization known as tonotopy or cochleotopy. This tonotopic organization is maintained throughout the auditory pathway, from the cochlea to the auditory cortex, helping to shape the brain's functional response to sound (Pulkki and Karjalainen, 2015; Roginska and Geluso, 2018).

2.1.2 Head-Related Transfer Function (HRTF)

Head-Related Transfer Functions (HRTFs) describe how sound travels from a point source in a space, to a specific point in the listener's ear canal (Møller et al., 1995; Li and Peissig, 2020). Differences in sound arriving at each ear—specifically in intensity

(loudness) and phase (timing)—are critical cues for sound localization. These differences, known as Interaural Intensity Differences (IIDs) and Interaural Time Differences (ITDs), occur because the head blocks some sound, reducing intensity at the ear farther from the source (IIDs), and because sound waves take longer to reach that ear (ITDs). These sound cues are captured using Finite Impulse Response (FIR) filters, which describe how sound is altered as it reaches the ear. In the time domain, these filters are called Head-Related Impulse Responses (HRIRs), while in the frequency domain, they are referred to as HRTFs. HRIRs capture how sound behaves over time as it reaches the ear from a specific location, while HRTFs describe how sound waves are filtered across different frequencies as they interact with the shape of the head, ears, and torso (Pulkki and Karjalainen, 2015; Li and Peissig, 2020; Roginska and Geluso, 2018; Cheng and Wakefield, 1999).

In applications like virtual reality (VR), augmented reality (AR), and mixed reality (MR), HRTFs are used to simulate six-degrees-of-freedom (6-DoF) binaural audio (Li and Peissig, 2020). The term "binaural" applies to scenarios where sound is sent to both ears, while "monaural" refers to cases where sound is directed to a single ear (Moore, 2012). By applying the HRTF to any sound, spatial characteristics can be imposed, making the sound appear to come from a specific location in space. When applied in real-time, these filters can simulate dynamic scenarios, such as sound source movement or changes in the listener's head position, enhancing the realism of virtual auditory experiences. To create an immersive experience in virtual space using headphones, the audio must be binauralization (Plinge et al., 2018). This process replicates how the human head, ears, and torso modify sound based on the source's direction and distance. Binauralization is achieved by convolving the audio signals with HRTFs, which correspond to the sound's relative position (Plinge et al., 2018). This technique helps the sound appear as though it is emanating from the surrounding environment enhancing the realism of the audio experience. In this research we employed the HRTFs while rendering the sounds within the 3D space as shown in Chapter 3 and Chapter 6.

2.1.3 How Are Sounds Modelled in Indoor Environments?

In 3D virtual environments, the propagation of sound within a room can be simulated and modeled using computer graphic ray tracing techniques (Savioja and Svensson, 2015; Potter et al., 2023). Drawing on the success of ray tracing in computer graphics, acousticians have effectively adapted these methods for the simulation of sound propagation in complex architectural spaces (Potter et al., 2023; Funkhouser et al., 2004). Existing approaches for acoustic modeling can be broadly divided into two categories: numerical solutions to the wave equation and geometrical acoustics methods, such as image source, ray tracing, and beam tracing techniques (Funkhouser et al., 2004; Savioja

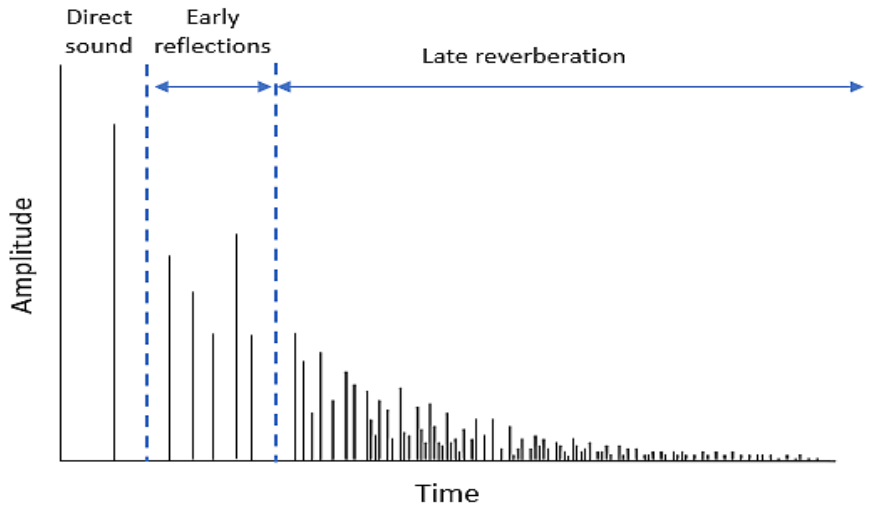

FIGURE 2.2: An illustrative representation of Room Impulse response (RIR).

and Svensson, 2015). Although numerical solutions based on the wave equation provide the highest level of accuracy, they are computationally intensive and unsuitable for interactive and dynamic applications (Funkhouser et al., 2004). Recent studies, such as (Borrel-Jensen, 2023), have employed deep learning to compute acoustic sound fields in simple rooms with varying shapes. In contrast, geometrical acoustics techniques, which assume that sound propagates as rays, offer faster computations and are widely used to model mid- and high-frequency behaviors in rooms where the sound wavelength is shorter than the surface dimensions and overall room size (Savioja and Svensson, 2015). However, at low frequencies, these methods become less accurate due to increased approximation errors, as wave phenomena play a more significant role (Savioja and Svensson, 2015). The image source method calculates specular reflection paths by reflecting the position of the audio source, as virtual sources across each polygonal surface in the environment (Funkhouser et al., 2004, 2002). Beam tracing method identifies propagation paths from a source by tracing beams that consist bundles of rays, covering the space of rays originating from the source within a 3D polyhedral environment (Funkhouser et al., 2004, 2002). Ray tracing, a commonly used approach in several simulation tools such as Unity [1], which is utilised in this research traces sound propagation paths by generating rays from the source position and following their interactions with the environment until a set of rays reaches the receiver (Funkhouser et al., 2004, 2002; Xiangyang et al., 2003). Unity, a VR gaming engine, is widely used in several VR applications, as elaborated in Section 2.5.

2.1.4 Room Impulse Response (RIR)

The impulse response signal serves as the fundamental source of information for understanding the audible characteristics of a room's sound field. The impulse response contains comprehensive details about the room's acoustics between a specific source and receiver (Rossing, 2007). Even when spatial information is omitted—assuming the source and receiver are modeled as point-like and omnidirectional—the impulse response still encapsulates the full room transfer function (Farina, 2007). This includes both time-domain effects such as echoes, discrete reflections, and the reverberant tail, as well as frequency-domain effects, including frequency response and frequency-dependent reverberation (Farina, 2007). The room impulse response (RIR) consists of direct sound, early reflections, and late reverberations (Hidaka et al., 2007; Stewart and Sandler, 2007; Remaggi et al., 2015). The direct sound is the initial sound received in a free-field environment, arriving without any reflections (Habets, 2007). Early reflections refer to the sounds that arrive to the receiver at first 50 or 80 milliseconds after the direct sound, these reflections resulting from sound waves reflecting off nearby surfaces such as walls, the ceiling, and furniture in the room (Hidaka et al., 2007; Habets, 2007; Fırat et al., 2022). Late reverberations are caused by sound reflections that arrive with longer delays after the initial direct sound (Habets, 2007). Figure 2.2 depicts a schematic representation of the RIR.

RIR Measurement. Several methods are used to capture the acoustic impulse response within a room, including the Maximum Length Sequence (MLS), which simulates the room's acoustics using a periodic pseudo-random signal with stochastic properties similar to pure white noise (Stan et al., 2002). The RIR is captured using circular cross-correlation between the measured output and the original MLS signal (Stan et al., 2002; Szöke et al., 2019). An alternative approach, known as the Inverse Repeated Sequence (IRS), consists of a sequence, where the first half corresponds to the MLS and the second half is the inverse of the MLS (Stan et al., 2002; Szöke et al., 2019). Moreover, the Time-Stretched Pulses method, which is also used to capture the RIR, relies on time expansion and compression of an impulsive signal to mitigate distortion peaks (Stan et al., 2002; Szöke et al., 2019). These methods, however, depend on the assumption that the system is linear and time-invariant, which can lead to distortion artifacts in the deconvolved impulse response if this condition is not satisfied (Stan et al., 2002). In (Farina, 2007), Farina introduced an enhanced approach to measuring the RIR for systems that are neither time-invariant nor linear by using the Exponential Sine Sweep (ESS) method. In ESS, frequencies vary exponentially over time, as illustrated in Equation 2.1, enabling effective capture of sound across different frequency bands. The frequency sweep starts at the lowest angular frequency, ω_1, and ends at the highest angular frequency, ω_2, over a duration of T seconds. However, the ESS method

[1] `https://unity.com/`(accessed in 2024)

effectively eliminates harmonic distortions, as they occur before the estimation of the linear impulse response, and it is also well-suited for quiet environments (Szöke et al., 2019).

$$x(t) = \sin\left[\frac{\omega_1 \cdot T}{\ln\left(\frac{\omega_2}{\omega_1}\right)} \cdot \left(e^{\frac{t}{T}\cdot \ln\left(\frac{\omega_2}{\omega_1}\right)} - 1\right)\right]. \tag{2.1}$$

To calculate the RIR captured by ESS, a deconvolution process is applied by convolving the output signal with the inverse filter of ESS, which is generated by temporally reversing the ESS signal (Farina, 2007). The deconvolution process is shown in Equation 2.2:

$$h(t) = y(t) * f(t), \tag{2.2}$$

where $h(t)$ is the impulse response, $y(t)$ is the recorded signal, and $f(t)$ is the inverse filter. In this research, in Chapter 3 and Chapter 6 we employed the ESS method to measure the RIR in the virtual space.

Acoustic Parameters. Since the RIR is captured and measured to represent the room's acoustics, many acoustic parameters can be evaluated to assess the quality of the room's acoustic geometry. According to ISO 3382-1:2009 (International Organization for Standardization, 2009), there are several key room acoustic parameters, including Early Decay Time (EDT), Reverberation Time 20 (T20), Reverberation Time 30 (T30), Clarity (C80), Definition (D50), Gravity Time (TS), and Sound Strength (G). Also, according to ISO 3382-1:2009 (International Organization for Standardization, 2009) the reverberation time RT60 can be measured using T20 or T30. 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). EDT is estimated using the slope of the decay curve, determined from a fit between 0 and -10 dB, and the decay time is calculated from this slope as the time required for a 60 dB decay (Barron, 1995; IoSR, 2024). 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). A well-known empirical formula to calculate RT60 is Sabine's formula (Rungta et al., 2016), shown in Equation 2.3:

$$\mathrm{RT}_{60} \approx 0.1611\,\mathrm{s\,m^{-1}}\,\frac{V}{Sa}. \tag{2.3}$$

This formula illustrates the relationship between the room volume V and the total absorption Sa, where S represents the surface area and a represents the absorption coefficient of the room surfaces. Another, more accurate way to estimate RT60 is to analyse the RIR and calculate the time it takes for the sound to decay by 60 dB, as specified in ISO 3382-1:2009 (International Organization for Standardization, 2009). In this research in Chapter 3 and Chapter 6, we measured EDT and RT60 according to this latter ISO 3382-1:2009 method (International Organization for Standardization, 2009). These two

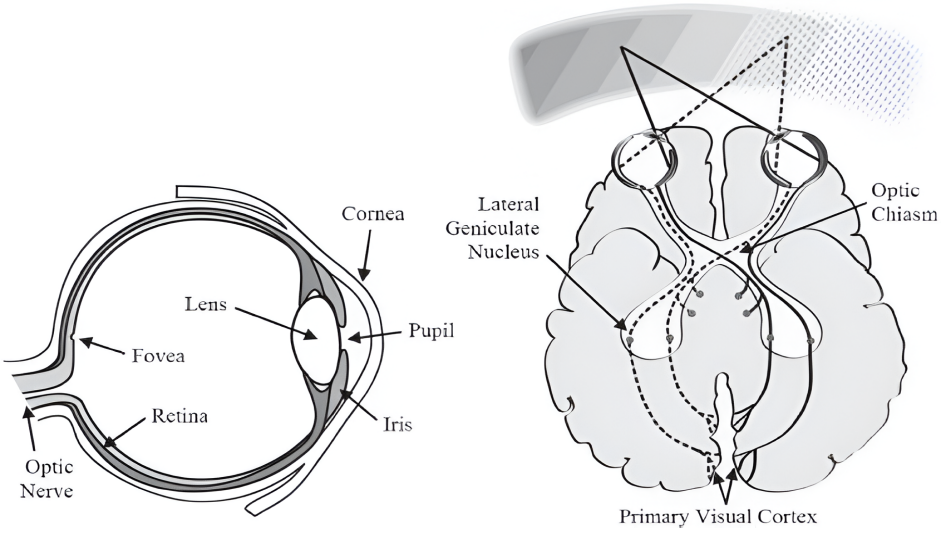

FIGURE 2.3: Human vision system. Image sourced from: (Bhowmik, 2017).

parameters are chosen due to their direct relationship with scene acoustic geometries that can be estimated using vision methods, and to enable comparison with similar studies in the literature.

2.2 3D Computer Vision

In this section, we present the background, along with the most relevant concepts and methods related to the vision aspect of our research.

2.2.1 How Do Humans See?

The human visual system functions as a complex imaging device, where light enters through the pupil and is focused onto the retina by the combined action of the cornea and lens, forming a two-dimensional (2D) image of the 3D world. This process is similar to the mechanism in modern camera sensors. The human visual system as shown in Figure 2.3, operates in a binocular manner, with each eye capturing a unique image from a slightly different location, resulting in two distinct perspectives of the 3D scene. The positional difference of corresponding points on the retinas, known as binocular disparity, is inversely related to the distance of the physical point from the viewer. Within the retina, light is converted into electrical signals that are transmitted to the brain's primary visual cortex via the optic nerve. The brain estimates depth by interpreting the binocular disparity between the images from each eye through a process called stereopsis (Bhowmik, 2017). Furthermore, human visual system employs monocular cues to understand the 3D spatial environment (Bhowmik, 2017). These monocular cues include texture variations, known object size, occlusions, and color,

etc. (Saxena et al., 2007; Reichelt et al., 2010). Therefore, the human visual system allows humans to easily perceive all objects even when they are partially occluded and have a natural ability to fill in the invisible parts (Chen et al., 2016; Ao et al., 2023).

2.2.2 Human Visual Perception in 3D Spatial Understanding

Humans are capable of perceiving the 3D shape from 2D retinal images with more stable and accurate way. This ability is supported by the integration of binocular and monocular cues in the human visual system, enabling an effective interpretation of spatial layout, depth, and object structure (Bhowmik, 2017). In the literature, several studies examined the contribution of these cues to 3D scene understanding and object recognition (Den Ouden et al., 2005; Bradshaw et al., 1998; Dövencioğlu et al., 2013; Li and Pizlo, 2011; Saarela and Landy, 2015). For example, (Den Ouden et al., 2005) demonstrated that color information improves binocular vision by improving depth perception through a more effective matching between images of the left and right eyes. Similarly, binocular disparity and motion parallax have been shown to be powerful for estimating the size and distance of objects (Bradshaw et al., 1998). When these cues are combined, perceptual accuracy increases, and the human visual system tends to integrate the outputs through cue averaging. The study by (Dövencioğlu et al., 2013) revealed that depth perception improves significantly when binocular disparity is paired with monocular shading cues, compared to when either cue is presented alone. Furthermore, the study in (Li and Pizlo, 2011) found that edge monocular cues when combined with binocular disparity, are more informative than other monocular cues, such as shading and texture. That study also emphasized the role of simplicity constraints, such as symmetry and planarity, which are fundamental for shaping perception under both monocular and binocular conditions. When such constraints are absent, even the presence of binocular disparity fails to ensure accurate 3D shape. In addition, monocular cues such as texture gradients, color, and luminance contribute to object recognition and are automatically integrated by the human visual system (Saarela and Landy, 2015).

In this research, we emulate these human cognitive mechanisms by integrating multiple visual cues into our 3D scene understanding framework. Specifically, we encode geometric depth using a flipped Truncated Signed Distance Function (F-TSDF), which enhances surface gradients and provides surface shape (more details in Section 2.2.4). We also incorporate additional monocular cues through RGB inputs such as color priors. As demonstrated in Chapter 4 and Chapter 5, this multi-cue approach significantly improves model performance, leading to more accurate 3D scene semantics and reconstruction.

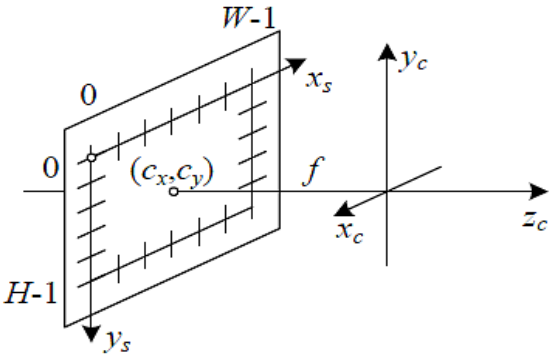

FIGURE 2.4: Camera intrinsics illustrating the focal length f and the principal point (c_x, c_y). The image width and height represented by W and F. Image sourced from: (Szeliski, 2022).

2.2.3 How Do Cameras Capture the 3D Space into 2D Image?

Perspective cameras follow the pinhole model projection (Sturm, 2021), and typically use camera intrinsic and extrinsic parameters to map points in 3D space onto a 2D image. Camera intrinsic parameters include the focal length and the principal point, as shown in Figure 2.4. The focal length represents the distance between the optical center (or camera center) and the image plane. The optical center is the origin of the coordinate system, while the principal point, which is the intersection of the optical axis and the image plane, is often set to the image center with $(c_x, c_y) = (W/2, H/2)$ (Zhang, 2021).

The camera intrinsic parameters are represented by the following matrix K (Szeliski, 2022):

$$K = \begin{bmatrix} f & 0 & c_x \\ 0 & f & c_y \\ 0 & 0 & 1 \end{bmatrix} \qquad K \in \mathbb{R}^{3\times3}.$$

The camera extrinsic parameters, or camera pose parameters, include the rotation and translation transformations that describe the relationship between the camera and the external world coordinates (Zhang, 2021). These are represented by a matrix $[R|t] \in \mathbb{R}^{3\times4}$. Therefore, the projection model is defined in Equation 2.4 as (Szeliski, 2022):

$$P = K[R|t], \tag{2.4}$$

some perspective cameras capture 3D depth information for each point, which is advantageous in various computer vision fields, such as 3D semantic scene completion (SSC) (Song et al., 2017; Wang et al., 2024a), room layout estimation (Song et al., 2015; Geiger and Wang, 2015), object detection and segmentation (Gupta et al., 2014, 2013),

FIGURE 2.5: Kinect camera by Microsoft. Image sourced from: Fablabs.io [2].

and 3D reconstruction (Handa et al., 2014; Zollhöfer et al., 2018). These cameras are known as depth sensors.

Depth Sensors. This type of sensor employs various technologies, such as structured-light and time-of-flight (ToF), both of which use infrared (IR) to capture objects in the scene. These cameras provide depth information for each pixel within a 2D image. Often, depth information is provided along with an RGB image, resulting in RGB-D data (Zollhöfer et al., 2018). For example, structured-light cameras project an IR pattern onto objects in the scene and estimate depth based on the perspective distortion of the pattern, which varies according to the object's depth (Zollhöfer et al., 2018). In contrast, ToF cameras measure the time it takes for IR light reflected off the object surface to travel back to a detector, thereby calculating the distance (Zollhöfer et al., 2018). Depth sensors are relatively inexpensive and are mostly suitable for indoor scenes, as their performance is affected by background light, and they lack precision in outdoor environments (Zollhöfer et al., 2018; Jimenez, 2021). Common commercial examples of depth-sensing cameras include the Kinect (Zhang, 2012) and RealSense [3]. Kinect v1 as shown in Figure 2.5, released in 2010, employed structured-light technology, while Kinect v2, released in 2014, used ToF technology. Alternatively, depth sensing technologies encompass non-perspective-based modalities, including Light Detection and Ranging (LiDAR) 360° sensors (Yang et al., 2023) and depth estimation based on stereo vision camera systems that employ distinct technological approaches (Laga et al., 2020; Szeliski, 2022). In this research, we process depth by Kinect camera, as it is used to collect the NYUv2 dataset (Silberman et al., 2012) as shown in Chapter 4 and Chapter 5.

Perspective cameras, however, suffer from a limited field of view (FoV) and cannot capture the full surroundings of a scene (Streckel and Koch, 2005; Zhang et al., 2016; da Silveira et al., 2022). Another type of camera with a larger FoV, capable of capturing full 360° surroundings, is known as a 360° camera (Gao et al., 2022). These cameras

[2] https://www.fablabs.io/machines/xbox-kinect (accessed in 2024)

[3] https://www.intel.com/content/www/us/en/architecture-and-technology/realsense-overview.html (accessed in 2024)

FIGURE 2.6: Ricoh Theta camera. Image sourced from: Ricoh Theta [4].

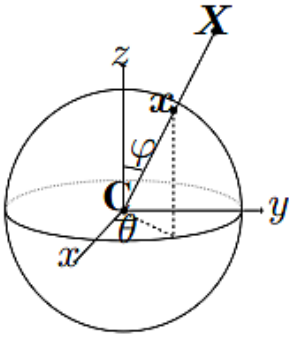

FIGURE 2.7: Illustration of the relationship between a point in 3D space and its corresponding point on a spherical image in the spherical camera model. Image adapted from (Akihiko et al., 2005).

are useful in computer vision applications, such as scene reconstruction (Dahnert et al., 2021; Li et al., 2024), as they provide a comprehensive view of the scene and its surroundings. Additionally, scenes captured by 360° cameras have expanded applications in VR, AR, and MR (da Silveira et al., 2022).

360° Sensors. In contrast to conventional pinhole cameras, 360° sensors capture the entire surrounding environment through spherical projection with stereo lenses (Li and Fukumori, 2005). The images produced by these sensors are commonly referred to as 360° images, full panoramas, spherical images, or omnidirectional images (da Silveira et al., 2022). These cameras are typically modeled as unit spheres, without intrinsic parameters, and are entirely defined by their extrinsic parameters (Krolla et al., 2014).

A spherical camera model is defined as "a camera center and a surface of a unit sphere whose center is the camera center" (Akihiko et al., 2005). The spherical camera center, $C \in \mathbb{R}^3$, is set as the origin of the world coordinate system. A spherical camera associates a ray from a point $X \in \mathbb{R}^3$ with the camera center C, as illustrated in Figure 2.7. The intersection between this ray and the spherical surface yields a point $x \in S^2$.

[4] https://www.ricoh360.com/theta/ (accessed in 2024)

Since x lies on a unit sphere, it can be expressed in spherical coordinates as $(r = 1, \theta, \varphi)$, where $\theta \in [0, 2\pi)$ and $\phi \in [0, \pi)$ (da Silveira et al., 2022). A common method for representing a 360° image on a 2D grid is equirectangular projection, which maps the latitude and longitude of the spherical image to horizontal and vertical grid coordinates (Coors et al., 2018). Another widely used technique to minimize distortions is cube-map projection, where the spherical image is represented as six equal perspective images (da Silveira et al., 2022; Huang et al., 2017; Kim and Hilton, 2015). This cube-map approach supports 2D computer vision tasks, such as material segmentation in perspective images, which can be extended to panoramic scenes, as detailed in Chapter 3.

Examples of such spherical cameras include the Samsung Gear 360 [5] and the Ricoh Theta [6], as shown in Figure 2.6. In this study, we utilised 360° scenes from the CVSSP dataset [7] , captured using the Ricoh Theta sensor, to create an immersive VR environment, as described in Chapter 3 and Chapter 6.

After describing methods for capturing 3D points in real-world space as 2D images, it is also essential to discuss techniques for representing the 3D geometry of the captured data within 2D images. This representation is critical for various computer vision applications that rely on reconstructing or analysing 3D data from 2D images.

2.2.4 What Methods Are Used to Represent 3D Geometry of Visual Data?

We will briefly describe some common 3D representations:

Point Clouds. A point cloud represents 3D data as X, Y, Z coordinates, usually capturing surface details (Wang and Kim, 2019). Point clouds are memory-efficient but lack a defined geometric structure, often requiring additional processing to construct a coherent 3D mesh, as depicted in Figure 2.8 (Mescheder et al., 2019). They are also susceptible to noise due to surface characteristics and sensor inaccuracies (Jimenez, 2021; Berger et al., 2017), and tend to be sparse, suffering from occlusions (Jimenez, 2021; Cheng et al., 2021; Berger et al., 2017). This type of data is commonly collected using depth cameras and LiDAR sensors.

Mesh. 3D meshes provide a compact representation of a scanned surface, composed of vertices, edges, and faces that together form polygonal shapes (Wang, 2024; Jimenez, 2021). Initially, vertices originate as point clouds, with edges defining the connections

[5] https://www.samsung.com/uk/support/mobile-devices/what-is-samsung-gear-360-camera/ (accessed in 2024)

[6] https://www.ricoh360.com/theta/ (accessed in 2024)

[7] http://3dkim.com/research/VR/index.html (accessed in 2024)

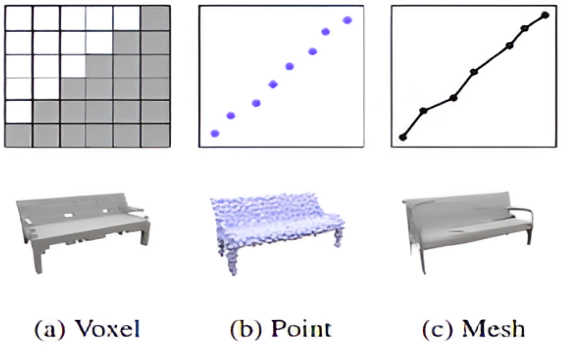

FIGURE 2.8: Various 3D representations segment the output space through different methods, illustrating differences in resolution and connectivity of the 3D data: (a) voxel representations divide the space spatially into a grid of units, (b) point-based representations focus on predicted individual points, and (c) mesh representations use vertices to define the structure. Image sourced from: (Mescheder et al., 2019).

between pairs of vertices, while faces represent areas bounded by edges connecting three or more vertices (Wang, 2024). Various methods, including but not limited to explicit geometry, can be employed to obtain 3D meshes (Berger et al., 2017).

Voxel Grids. Voxel grids provide a discrete volumetric representation of 3D space, where each cubic unit is called a voxel (Roldao et al., 2022; Mescheder et al., 2019), as shown in Figure 2.8. They are often used to store occupancy information derived from point clouds (Roldao et al., 2022). For example, a voxel grid can store depth map points for indoor scenes in a binary format, indicating whether each voxel represents a visible surface. This is achieved by defining the voxel size and transforming pixel depth values to 3D coordinates within the voxel grid using camera parameters for precise representation (Dourado Neto, 2024). However, voxel grid representation consumes substantial memory, as it stores data for both occupied and free spaces (Mescheder et al., 2019; Roldao et al., 2022). A unique advantage of voxel representation, not available in other methods, is its ability to capture internal geometry, making it useful for representing complex 3D structures, such as the indoor scenes in the NYUv2 dataset used in this research (Silberman et al., 2012). Voxel grids can also store additional information, such as color and density (Wang, 2024; Dourado Neto, 2024).

Implicit Representations. Techniques such as the Signed Distance Function (SDF) provide a compact and continuous representation of surfaces derived from 2D images (Jiang et al., 2020; Wang, 2024; Fatima, 2024). Compared to meshes, SDFs offer the advantage of representing arbitrary topologies. Unlike point clouds, which are inherently sparse, SDFs represent watertight surfaces (Jiang et al., 2020). An SDF encodes an object's geometry by specifying the distance from any point in space to the closest surface

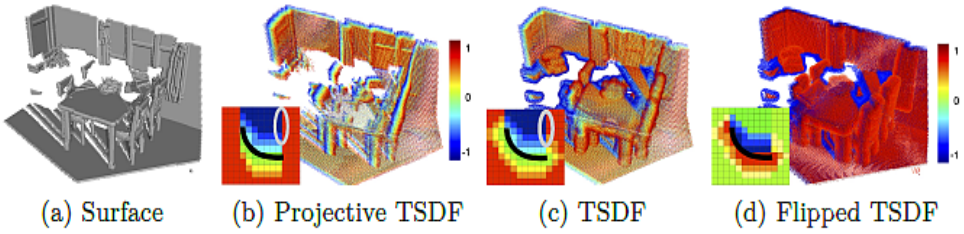

FIGURE 2.9: Various encoding methods for surface (a). The projective TSDF (b) is computed with respect to the camera angle and is therefore view-dependent. The accurate TSDF (c) shows reduced view dependency but has strong gradients in empty space near the occlusion boundary. In comparison, the flipped TSDF (d) displays the highest gradient close to the surface. Image sourced from: (Song et al., 2017).

of the object. The sign of the distance indicates whether the point lies inside (negative) or outside (positive) the object (Chen et al., 2022; Wang, 2024; Fatima, 2024).

The projective Truncated Signed Distance Function (TSDF) was introduced in 2011 with advancements in 3D reconstruction (Newcombe et al., 2011). The authors developed a real-time reconstruction system for indoor scenes using a Kinect sensor, where all depth data streamed from the Kinect is fused into a single global implicit surface model of the observed scene. The TSDF is a variant of the SDF that restricts distance values to a specified threshold around the surface, thereby reducing both computational load and memory usage (Newcombe et al., 2011). Building on this, (Song et al., 2017) proposed a modification to the projective TSDF used in (Newcombe et al., 2011), where each voxel records the Euclidean distance to the nearest surface, with the sign indicating whether the voxel is located in free space or within an occluded region. The approach in (Song et al., 2017) eliminates view dependency by calculating distances to the nearest point across the entire observed surface, rather than limiting calculations to the line of sight of the camera. They further addressed the issue of strong gradients in empty space near occlusion boundaries by introducing a flipped TSDF (F-TSDF), as illustrated in Figure 2.9. The F-TSDF is calculated as:

$$F\text{-TSDF} = \text{sign}(TSDF) \cdot (TSDF_{\text{max}} - |TSDF|) \tag{2.5}$$

The sign in Equation 2.5 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. In this research, we adopt F-TSDF method (Song et al., 2017) for the implicit representation of 2D images in the NYU datasets (Silberman et al., 2012; Firman et al., 2016).

2.2.4.1 Can We Learn 3D Representations Using Deep Learning from a Single View of an Indoor Scene?

In recent literature, point clouds have been widely used for 3D semantic segmentation of indoor scenes, often employing convolutional neural networks (CNNs) (Hua et al., 2018; Qi et al., 2017; Tchapmi et al., 2017; Guo et al., 2024; Xu et al., 2024; Hu et al., 2023). Some approaches have also utilised Recurrent Neural Networks (RNNs) for this purpose (Huang et al., 2018). Additionally, point clouds have been applied to 3D reconstruction and completion tasks (Wen and Cho, 2023; Li et al., 2024). Meshes have also been used for 3D reconstruction in studies such as (Nie et al., 2020; Hu et al., 2022; Fang et al., 2021). However, these methods are limited to learning the 3D representation of surface geometry alone, relying on explicit 3D representation input.

However, shape completion in complex scenes is a key research area in computer vision that focuses on geometry processing. In contrast to 2D image completion, which relies on the availability of extremely large numbers of similar images or the assumption that the necessary structure for completion is present in the input data, 3D completion algorithms are concerned with estimating full 3D occupancy. The goal of 2D image completion is typically to produce a visually plausible output rather than to predict the unobserved ground truth accurately (Firman et al., 2016). Completion algorithms initially used interpolation or energy minimization methods to complete small missing regions (Roldao et al., 2022). Completing partial input by reasoning from geometric cues using plane fitting or object symmetry often fails when the missing regions are large (Song et al., 2017). On the other hand, fitting predefined 3D mesh models to the input depth map is a common approach for inferring the full geometry and semantic labelling of a scene (Gupta et al., 2015; Song and Xiao, 2014; Geiger and Wang, 2015). The quality, quantity, and diversity of 3D models that are accessible for retrieval are restricting the prediction quality. By applying this method, it is observed that objects which the current models cannot describe are often overlooked. Alternately, if the 3D model dataset is enormous enough to hold all observations, retrieval and alignment difficulties must be resolved (Song et al., 2017). To solve these problems, (Lin et al., 2013; Song and Xiao, 2016; Jiang and Xiao, 2013) utilised 3D primitives such as cuboids to approximately define the complete 3D geometry of detected objects instead of using complete 3D mesh models as reference. This method has the obvious drawback of only being able to supply rough shape information, which is inappropriate for geometry completion (Firman et al., 2016). Studies employing learning-based methods have demonstrated that these approaches are more adaptable and efficient compared to the previous ones. Typically, it employed deep neural networks, which have a quick inference speed and superior robustness, to infer the invisible area (Chen et al., 2020). (Dai et al., 2017b) proposed a 3D-Encoder-Predictor Network that first encodes the known and unknown space to obtain a comparatively low-resolution prediction, and then connects this intermediate result with 3D geometry from a shape database. Using raw point clouds directly,

without making any structural assumptions about the underlying shape, is what (Yuan et al., 2018) introduced as an end-to-end solution. (Stutz and Geiger, 2020) provided a technique for 3D shape completion based on weakly supervised learning. They trained a shape prior on synthetic data and learned maximum likelihood fitting using CNN. (Zhou et al., 2021) proposed probabilistic generative modelling of 3D shapes using Point-Voxel Diffusion (PVD). PVD combines hybrid point-voxel representations of 3D forms with denoising diffusion models. It trained by maximising a variational lower bound to the conditional likelihood function and it could be seen as a series of denoising stages that reverse the diffusion process from observed point cloud data to Gaussian noise. These techniques concentrate on reconstructing 3D shapes from the incomplete input of a single object, which makes it difficult to be extended to partial scenes with multiple objects approximated at the semantic level.

A more recent problem in computer vision, termed semantic scene completion (SSC), was introduced by (Song et al., 2017). SSC aims to learn 3D representations that encompass not only visible surfaces but also challenging regions in occluded areas, from a single partial view within a 2D image. This approach leverages voxel grids, as discussed in Section 2.2.4, to capture the internal geometry of occluded regions. Additionally, SSC utilises an implicit representation based on the TSDF to encode 2D images, enabling the capture of 3D shapes and structures in a more comprehensive manner. More elaboration on SSC will be provided in Section 2.3.

In this research, we propose two models contributing to SSC: a model that uses depth-only input, detailed in Chapter 4, and an enhanced model with dual input (RGB and depth), as described in Chapter 5.

2.3 Semantic Scene Completion (SSC) for Perspective Indoor Views

SSC is a relatively recent research field that began with the work of Song et al. in 2017, who introduced SSCNet, the first deep neural network designed specifically for SSC (Song et al., 2017). The SSC task involves simultaneously predicting volumetric occupancy and object categories at the voxel level from a partial view. Thus, SSC encompasses two closely related tasks: scene completion and semantic scene completion (Song et al., 2017). Additionally, they provided the SUNCG synthetic dataset including 3D densely annotated models for indoor scenes, which has a significant number of scenes that are inaccessible due to violations of intellectual property rights [8].

[8] https://https://futurism.com/tech-suing-facebook-princeton-data/ (accessed in 2023)

TABLE 2.1: Indoor scenes datasets.'# Frames' means the number of images in the dataset with splitting if provided.

Dataset	Scenes Type	Input → Ground Truth	# Frames
NYUv2 (Silberman et al., 2012)	Real data	RGB-D → Voxel	795/654
SceneNN (Hua et al., 2016)	Real data	RGB-D → Mesh	100
ScanNet (Dai et al., 2017a)	Real data	RGB-D → Mesh	1201/312
CompleteScanNet (Wu et al., 2020)	Real data	RGB-D → Mesh/Voxel	45448/11238
ScanNet++ (Yeshwanth et al., 2023)	Real data	RGB-D → Mesh	360/50/50
NYUCAD (Firman et al., 2016)	Synthetic data	RGB-D → Voxel	795/654
SceneNet (Handa et al., 2016)	Synthetic data	RGB-D → Mesh	57
PCSSC-Net (Zhang and Wonka, 2021)	Synthetic data	RGB-D → Points	1520/392

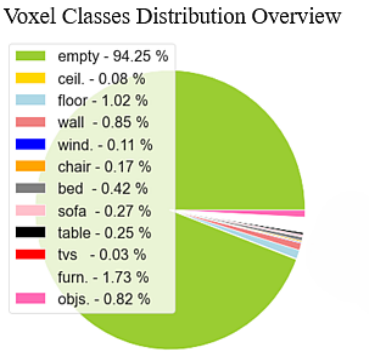

FIGURE 2.10: An illustration of 3D data distribution within NUY training set.

Song et al. observed that occupancy patterns within scenes are strongly correlated with the semantic labels of objects. Consequently, predicting voxel occupancy and identifying object semantics are highly interdependent. Recognizing an object's identity aids in predicting its likely position within a scene. For instance, if the top of a chair is visible behind a table, it may be inferred that the chair's seat and legs are present as well. Conversely, identifying an object's occupancy pattern can facilitate the classification of its semantic category (Song et al., 2017).

As SSC integrates both scene completion and semantics, the following sections will provide an overview of indoor scene datasets relevant to the SSC problem, as well introduce existing SSC architectures, discuss loss functions designed to learn scene geometries, and outline evaluation metrics, along with training and evaluation schemes.

2.3.1 Indoor Scenes Datasets

There are several indoor datasets available in the literature, as shown in Table 2.1. However, researchers have been using NYU datasets (NYUv2 (Silberman et al., 2012), and NYUCAD (Firman et al., 2016)) since the time when the SUNCG dataset was available, and continue to use them till now. The NYU datasets are the best option for the defined research problem compared to other indoor datasets due to their real and synthetic data, dense annotations for volumetric representation, a wide range of scenes, and dataset size. Compared to NYUv2 and NYUCAD, datasets such as SceneNN (Hua et al., 2016), ScanNet++ (Yeshwanth et al., 2023), and SceneNet (Handa et al., 2016) are considerably smaller. The PCSSC-Net dataset (Zhang and Wonka, 2021), while synthetic and represented with point clouds, lacks sufficient local connectivity, which limits its utility in completion tasks. The ScanNet dataset (Dai et al., 2017a), primarily designed for object classification, retrieval, and surface-level semantic voxel labeling, does not offer the dense annotations required for volumetric tasks. The CompleteScanNet dataset (Wu et al., 2020), which incorporates ShapeNet (Wu et al., 2015) and Scan2CAD (Huang et al., 2021), is subject to ground truth inconsistencies, with missing data in scene elements such as walls, floors, and ceilings.

Given these considerations, this research utilises the NYU datasets (NYUv2 (Silberman et al., 2012) and NYUCAD (Firman et al., 2016)) for the semantic scene completion (SSC) tasks detailed in Chapter 4 and Chapter 5. However, the NYUv2 dataset comprises 1449 realistic RGB-D indoor scenes captured with a Kinect sensor at a 640×480 resolution. The scenes are challenging and complex. For this dataset the voxelized 3D ground truth data derived from the annotations in (Guo et al., 2015), with object categories mapped according to (Handa et al., 2016). The dataset is divided into 795 training instances and 654 testing instances. However, as noted by (Song et al., 2017), the NYUv2 dataset suffers from occasional misalignments between depth images and their 3D labels. The NYUCAD (Firman et al., 2016) addressed this issue as it provides depth maps rendered directly from ground truth annotations, thus eliminating alignment issues. The NYU 3D voxelised data represent a significant imbalance in 3D space, with approximately 95% of voxels unoccupied and only 5% occupied. Figure 2.10 shows this imbalance within the NYU 3D data training set. In this research, we contribute to address this challenge in Chapter 4.

2.3.2 SSC Architecture Designs and Input Modalities

The design of SSC architectures is closely tied to the type of input data, including 3D geometry representations derived from depth maps using 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).

Volume Networks. 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 normalized range of -1 to 1 (Song et al., 2017; Garbade et al., 2019). Many studies, such as (Song et al., 2017; Zhang et al., 2018b,a, 2019; Dourado et al., 2021), use F-TSDF to provide steeper gradients at surface boundaries.

A common practice in volume-based networks is the use of encoder-decoder architectures with skip connections to retain contextual information (Roldao et al., 2022). Some studies such in (Zhang et al., 2018b) extended SSCNet by incorporating a dense Conditional Random Field (CRF) on the output, combining unary potentials (SSCNet output) with pairwise potentials by TSDF derived from depth geometries. Similarly, (Zhang et al., 2018a) proposed spatial group convolutions that operate orthogonally on spatial dimensions rather than feature channels.

Further advancements include CCPNet (Zhang et al., 2019), which introduced a Guided Residual Refinement (GRR) module. This module uses hyperbolic tangent function-based connections to amplify fused features and restore the fine structure of objects. Additionally, EdgeNet (Dourado et al., 2021) leveraged the UNet architecture and incorporating edge information from RGB data into the voxelized representation within data preprocessing step, guiding the model's learning process.

View-Volume Networks. 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., 2020b, 2019a, 2020a; Zhong and Zeng, 2020). One of the earliest methods to incorporate RGB features with depth data is (Liu et al., 2018) in SSC domain, which proposed a reprojection layer based on camera intrinsic parameters. (Li et al., 2019a) introduced the lightweight Dimensional Decomposition Residual network (DDR) designed for 3D dense prediction with the advantage of reducing the network parameters. The DDR breaks down the standard 3D convolution into three sequential one-dimensional layers along three orthogonal axes. (Li et al., 2020b) introduced a multi-modal fusion architecture that employs 2D semantic segmentation to guide 3D features taking the advantage of residual attention block (RAB) that combines both channel and spatial attention modules with DDR block. (Li et al., 2020a) proposed anisotropic convolutions by decomposing 3D convolutions into three sequential one-dimensional convolutions, where the kernel size of each one-dimensional convolution

is adaptively determined. Similarly, the study in (Li et al., 2023) proposed Planar Convolution Residual (PCR) block, a variant of the Dimensional Decomposition Residual (DDR) block in (Li et al., 2019a). The PCR is based on planar convolutions with kernel dimensions where one of the three sizes is 1. Additionally, they proposed attention module for capturing the global context from the front surface to the rear occluded areas. Employing point cloud inputs, as explored by Zhong et al. (Zhong and Zeng, 2020), present challenges such as sparsity and a lack of local structural detail, which are typically not present in 3D occupancy grids.

Other works, such as (Cao and de Charette, 2022; Wang et al., 2024b; 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 or shape information. (Cao and de Charette, 2022) introduced a 3D context relation prior to enhance spatio-semantic awareness by learning semantic scene-wise relation maps using a 3D UNet bottleneck. (Wang et al., 2024b) proposed a method to first predict depth maps, which are then fused with RGB features. Additionally, they designed a confidence-aware 2D-3D projection mechanism that compares feature certainty with a trainable latent variable to refine the 3D projection positions. Building on (Cao and de Charette, 2022), (Yao et al., 2023) proposed a method to extend 2D feature maps into 3D space by gradually reconstructing the depth dimension through deconvolution operations.

Hybrid Networks. 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., 2019b; Chen et al., 2020; Cai et al., 2021; Wang et al., 2022; Dourado et al., 2022; Wang et al., 2023). (Chen et al., 2020) proposed the 3D Sketch model, which employs a Sketch Hallucination Module that utilises the semi-supervised structural prior property of a Conditional Variational Autoencoder (CVAE) to guide full 3D sketch inference from partial observations. SISNet (Cai et al., 2021) proposed a method for learning 2D instance semantics through iterative scene-to-instance and instance-to-scene semantic completion. The approach reconstructs instances using a backbone and proposal module to determine their locations, sizes, and categories, followed by voxelizing them at higher resolution to recover detailed 3D shapes. FFNet (Wang et al., 2022), building on (Chen et al., 2020), integrates RGB and depth data by performing 2D feature correlation in the frequency domain. SPAwN (Dourado et al., 2022) proposed an encoder-decoder architecture and incorporating surface normals alongside geometric and RGB features. Similarly, (Wang et al., 2023) introduced a knowledge distillation-based model called Cleaner Self (CleanerS). This model comprises two networks: a teacher network trained on clean TSDF data from the NYUCAD dataset with rendered depth maps, and a student network trained on noisy TSDF data from the Kinect sensor within the

NYUv2 dataset. Both networks include additional RGB features, with the teacher network providing intermediate supervision to the student network for both structure and semantics. (Chen et al., 2019) explored adversarial training with a conditional generative adversarial network (GAN), though it achieved suboptimal results compared to the recent study by (Wang et al., 2024a), which applied adversarial training with tailored guidance. This approach enhanced the generator's ability to ensure visual fidelity by addressing both geometric completeness and semantic accuracy. Additionally, (Lin et al., 2023) extended the work in (Li et al., 2019a) by incorporating multi-head attention and multi-scale feature fusion. (Fu et al., 2023) combined point clouds with voxel representations, introducing a Surface-Attention module to guide the features of voxels near the surface.

However, 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., 2016a; Russakovsky et al., 2015). Research by (Garbade et al., 2019; Wang et al., 2022) utilised the pre-trained Deeplab v3+ (Chen et al., 2018) on the ADE20K dataset (ADE). A recent study by Wang et al. (Wang et al., 2023) employed the Segformer (Xie et al., 2021), initialized 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. In this research, we propose two models: a volume-based model and a hybrid model that simultaneously train on two distinct representations of the scene with F-TSDF and RGB inputs. Our baseline design is inspired by the UNet architecture in (Dourado et al., 2021). Additionally, 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. More details are discussed in Chapter 4 and Chapter 5.

2.3.3 Loss Function Designs for SSC Modeling

The design of loss functions is often closely tied to the model architecture, as seen in approaches employing adversarial training. For instance, (Chen et al., 2019) utilised an adversarial training objective function customized to address the challenges of the

SSC problem. Similarly, (Chen et al., 2020) proposed a combined loss comprising cross-entropy, Kullback-Leibler divergence, and semantic losses to provide a self-supervised semantic signal. Their approach incorporated a Sketch Hallucination Module, leveraging the semi-supervised structural prior property of a Conditional Variational Autoencoder (CVAE) to guide full 3D sketch inference from partial observations. They employed the Kullback-Leibler divergence loss to encourage the model to generate a distribution aligned with a predefined prior distribution. Other methodologies incorporate point clouds along with voxels such in (Tang et al., 2022; Fu et al., 2023) employed additional loss for points semantic features supervision.

It has been observed, few studies have proposed loss functions that explicitly address the data imbalance problem through weighting techniques. Studies in (Song et al., 2017; Zhang et al., 2019) used a voxel-wise cross entropy with $\{0, 1\}$ weights for empty and occupied voxels with undersapmling the occluded empty voxels to tackle the imbalance between occupied and empty voxels. PALNet (Li et al., 2019b) introduced a Position-Aware Loss (PA) that uses local geometric information to calculate voxels' weights, encouraging the identification of voxels with rich details. They employed Local Geometric Anisotropy (LGA) to assess the differences between a voxel and its neighboring voxels. Tang et al. (Tang et al., 2022) introduced a weighted cross-entropy loss by giving 0 weight for visible empty and 1 otherwise. This loss is combined with a semantic-aware loss that supervises the pairwise similarity in the generated point clouds. Dourado et al. (Dourado et al., 2021, 2022) randomly select the occluded voxels to achieve a balance between occupied and empty voxels. Additionally, (Dourado et al., 2022) assigned higher weights $(w = 2)$ to rare categories such as TV.

However, the weighting methods mentioned above largely overlook the issue of category imbalance within datasets. In this research, we propose a novel re-weighting method that addresses the pronounced imbalance between occupied and non-occupied voxels while also considering the category imbalances within the NYU datasets (Silberman et al., 2012; Firman et al., 2016), as detailed in Chapter 4.

2.3.4 Evaluation Metrics

The proposed evaluation metrics for SSC in indoor scenes with voxelized data include Precision, Recall, and Intersection over Union (IoU) for evaluating the scene completion task, as well as mean Intersection over Union (mIoU) (Everingham et al., 2015; Song et al., 2017) for assessing the semantic completion task, excluding empty space, as depicted in Equation 2.6:

$$mIoU = \frac{1}{C} \sum_{c=1}^{C} \frac{TP_c}{TP_c + FP_c + FN_c}. \tag{2.6}$$

Equation 2.6 calculates the mean Intersection over Union (*mIoU*) by using true positives (*TP*), false positives (*FP*), and false negatives (*FN*) for each class (*C*). In the scene completion task, the IoU metric is based on the formula in 2.6 but does not incorporate semantic class labels. Specifically, voxels in occluded regions are binary classified: non-empty voxels are assigned a '1', while empty voxels are assigned a '0'. We observe a lack of standardized methods for selecting the scene completion area for IoU evaluation, which complicates result comparison across studies. For instance, SATNet (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 unoccupied occluded voxels. Other studies, such as PALNet (Li et al., 2019b), DDRNet (Li et al., 2019a), and AICNet (Li et al., 2020a), include all occupied voxels in the scene, combining visible surfaces with occluded regions for scene completion evaluation.

Additionally, many studies fail to provide detailed descriptions, methods, or code for selecting occluded regions, which further complicates reproducibility and comparison. As a result, mIoU is considered to better reflect real model performance and is more critical than IoU for comparing state-of-the-art (SOTA) models (Liu et al., 2024; Li et al., 2020a). In this research, we present our results using both IoU and mIoU, along with Precision and Recall. For IoU we evaluated all the occluded occupied voxels and following the approach in (Liu et al., 2018), we consider the resampled empty occluded voxels.

2.3.5 Training and Validation Scheme

Most studies in indoor SSC literature adopt the hold-out procedure, which involves dividing the dataset into two splits: a training set and a hold-out set for evaluating model performance. However, it is often unclear whether these studies also incorporate a validation split during training to monitor the model's fitting behavior. Without a validation split, detecting underfitting or overfitting becomes challenging, yet these are critical for assessing model performance. Limited studies used validation set, a recent study (Wang et al., 2024a) reported randomly sampling 100 instances from the training set to use as a validation set. Another limitation observed, that none of the SSC studies in the literature have, to our knowledge, reported performance uncertainty by providing average results with standard deviations. Instead, some studies, such as (Zhang et al., 2018a), explicitly mentioned that they reporting only their highest achieved scores. Reporting average results with standard deviations is an essential step toward building generalised and unbiased results. Furthermore, we observe minimal variation in reported scores across different models, making it challenging to compare performance effectively without standard deviations calculated over multiple runs.

In machine learning literature, K-fold cross-validation is a widely recognized technique for measuring model performance (Nti et al., 2021) and assessing generalisation error (Anguita et al., 2005; Blum et al., 1999). This technique involves randomly partitioning the dataset into k_f equal-sized subsets (folds). Each fold is used as validation data while the remaining k_f - 1 folds are used for training. The process is repeated k_f times, and the results from all folds are averaged to yield a single performance estimate for the model (Stone, 1974; Wong, 2015; Rodriguez et al., 2009). In this research, we employed K-fold cross-validation to address the gap in SSC model evaluation by assessing performance uncertainty. We evaluated model performance using a single depth input in Chapter 4 and examined multiple inputs with various fusion strategies of incorporating RGB features in Chapter 5.

2.4 3D Indoor Scene Reconstruction and Completion from 360° Camera Views

Significant progress has been made in constructing 3D models of indoor scenes from perspective camera views, as discussed in Section 2.3. Methods leveraging single-depth inputs (Song et al., 2017; Zhang et al., 2019, 2018a) or solely RGB data (Cao and de Charette, 2022), as well as approaches combining RGB-D inputs (Wang et al., 2023; Dourado et al., 2022; Chen et al., 2020), have demonstrated the ability to generate semantically labeled 3D voxel structures. These structures include both visible and occluded regions and are constructed using convolutional neural networks (CNNs) trained jointly for semantic segmentation and scene completion. However, these methods are constrained by their limited input modalities and partial scene coverage, making them inadequate for applications requiring fully immersive VR environments.

In contrast, relatively few studies have addressed 3D reconstruction from 360° inputs, which provide a complete scene view. For instance, (Li et al., 2024) employs CNNs for surface reconstruction, but this approach does not extend to generating annotated 3D models with semantic segmentation. Similarly, the work in (Kim et al., 2022), demonstrates densely annotated 3D models using depth-only 360° inputs. Nevertheless, a gap remains in developing frameworks that integrate RGB data for fully immersive and annotated 3D reconstructions.

In this research, we extend the inference capabilities of the pre-trained MDBNet model introduced in Chapter 5. Originally trained on densely annotated datasets of indoor perspective scenes, MDBNet is adapted to process 360° RGB-D inputs, enabling the generation of comprehensive 3D models suitable for immersive VR environments, as detailed in Chapter 6. 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.

2.5 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 voxelized 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 depend not only on room shape and size but also 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 [9] with sound spatialisation plug-ins: Google Resonance Audio [10] in (Kim et al., 2019) and Steam Audio [11] 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 [12], CryEngine [13], AppGameKit VR [14], ApertusVR [15], and Urho3D [16]. 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; Vuorinen). Unity is particularly noted for its lightweight build and ease of use compared to Unreal Engine (Sabir et al.; 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 HRTFs to improve localization accuracy. Similarly, (Røsvik, 2024) developed a VR orchestral concert experience using Unity combined with the Oculus Spatializer Native toolkit for audio specialization.

In our research, we conduct preliminary experiments using EdgeNet360 as described in (Kim et al., 2022), with a 360° RGB only input, to construct a VR space, as demonstrated in Chapter 3. The VR space reconstruction is enhanced in Chapter 6, by leveraging the SSC model introduced in Chapter 5. This approach enables the creation of more comprehensive and immersive VR spaces, addressing existing limitations in the field.

2.6 Summary

In this chapter, we introduce key concepts related to the data used in this research. Additionally, we review existing methods for constructing VR spaces from a single image and explore VR gaming engines used to create virtual environments with spatial sound rendering. These methods and tools are linked to our preliminary results in Chapter 3, which serve as the motivation for this research. Our final results, along with a comprehensive evaluation and comparison with existing work in the literature, are presented in Chapter 6.

We also detail the deep learning architectures employed for SSC in indoor views, as well as the datasets and evaluation methods used for a thorough comparison with our

[9] https://unity.com/(accessed in 2024)
[10] https://resonance-audio.github.io/resonance-audio/(accessed in 2024)
[11] https://valvesoftware.github.io/steam-audio/(accessed in 2024)
[12] https://www.unrealengine.com/en-US(accessed in 2024)
[13] https://www.cryengine.com/(accessed in 2024)
[14] https://www.appgamekit.com/dlc/vr(accessed in 2024)
[15] https://apertusvr.org/(accessed in 2024)
[16] https://urho3d.io/(accessed in 2024)

proposed models in Chapter 4 and Chapter 5. The SSC model proposed in Chapter 5 forms the foundation for constructing 3D voxel models, as demonstrated in Chapter 6.

The following chapter delves into the motivation behind this research, focusing on the construction of VR spaces from single 360° images.

Chapter 3

Audio-Visual Scenes Generation in Virtual Space Using Single 360° Scene (A Preliminary Work)

3.1 Motivation and Contribution

This chapter introduces preliminary work on constructing an audio-visual space from a single 360° input. In this work, we propose a pipeline that integrates scene semantics and material properties from a single RGB 360° input, enabling realistic acoustic rendering for an enhanced auditory experience. It addresses the first research question, RQ1, and lays the foundation for the methodological choices in this research.

The motivation for this work stems from the need to simplify the traditionally resource-intensive process of audio-visual scene generation. Classical methods for estimating room acoustic properties require physical setups involving microphones and loudspeakers, which can be time-consuming and resource-demanding (Kon and Koike, 2018). Instead, this work introduces a computer vision-based approach, where a single 360° camera image can be used to reconstruct the 3D geometry of the scene and estimate room acoustic parameters via synthesised room impulse responses (RIRs).

This approach leverages the fact that accurate 3D geometry with material properties is a key factor in simulating realistic room acoustics, as 3D models can closely mimic real-world acoustic behavior (Hulusic et al., 2012). By estimating both the 3D scene semantics and material properties directly from visual input, the proposed pipeline significantly reduces the need for additional hardware setups. Moreover, the proposed pipeline not only offers a practical and efficient way to generate VR spaces but also

opens new possibilities for dynamically estimating room acoustics in real-time, offering a scalable solution for applications in entertainment, education, and architectural design.

Previous works have explored audio-visual scene generation but differ in methodology. For instance, the study in (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 estimated by stereo cameras. The resulting 3D point cloud is 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 another study, (Kim et al., 2022) employed EdgeNet (Dourado et al., 2021), a semantic scene completion (SSC) deep learning model, to infer scene semantics directly within 3D space. Once the 3D models are built in these studies, sound is rendered within the scenes to create a coherent and immersive experience by integrating both audio and visuals.

In contrast to previous works, which rely on multiple cameras for depth estimation (Kim et al., 2020a, 2021), this research proposes a more efficient solution using only a single 360° image. This eliminates challenges related to camera synchronization and alignment. Additionally, the inclusion of material segmentation (derived from both local and global features) ensures a more accurate representation of surface properties, as demonstrated in (Heng et al., 2022), which are crucial for acoustic modeling. The proposed pipeline integrates recent advancements in computer vision to streamline the process of VR scene generation, enhancing immersive audio-visual experiences.

In this chapter we contributed by:

- Extending material estimation inference from perspective RGB images to 360° RGB inputs through cubic projection, using a decomposition and composition process to align the proposed pipeline components effectively.
- Constructing a VR space using the Unity platform integrated with the Steam Audio plug-in, followed by evaluating room acoustic parameters in the virtual environment.

3.2 Proposed System

3.2.1 Overview

This pipeline aims to construct 3D virtual space from a single 360° photo of an indoor scene with sound rendering capabilities. As shown in Figure 3.1, the process begins with capturing a complete indoor scene using an off-the-shelf 360° camera. The

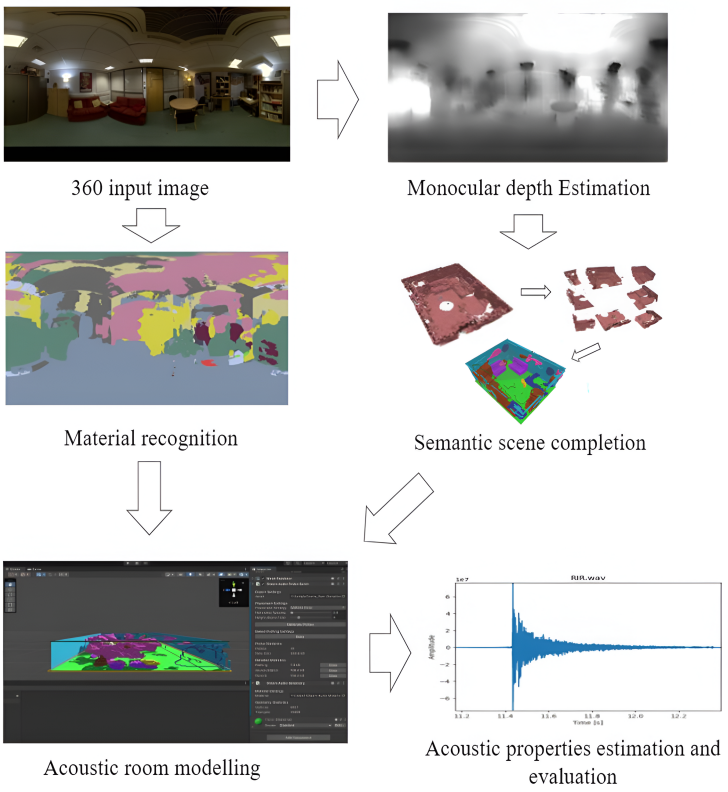

FIGURE 3.1: End-to-end system structure: a single 360° image input to estimate monocular depth. Both materials recognition and 3D model reconstruction are processed in parallel. Results are integrated into Unity platform for complete 3D scene with materials labels to generate plausible sounds in VR space.

pipeline integrates output from three main components: monocular depth estimation using a U-Net model to estimate the depth of the scene (Wu et al., 2021). EdgeNet360 (Kim et al., 2022), an extension of EdgeNet (Dourado et al., 2021) designed for predicting comprehensive 3D geometrical structure by completing the scene's invisible parts and predicting semantic labels. Material estimation model based on (Heng et al., 2022) for predicting materials in a given RGB input.

The outputs from these models are integrated into the Unity platform for both geometry rendering and sound reproduction. By combining the 3D geometry with material properties, the system allows for the evaluation of room impulse response (RIR) acoustic parameters, such as early decay time (EDT) (Barron, 1995) and reverberation time (RT60) (Rungta et al., 2016). These acoustic measures are used to compare the reproduced room model with the actual recorded sound in the space. The following sections detail the methods employed within the pipeline.

3.2.2 Monocular Depth Estimation

A U-Net shape encoder-decoder model for a single 360° image depth estimation is used based on (Wu et al., 2021). The encoder uses ResNet50 (He et al., 2016a) as its backbone, while the decoder consists of two convolution layers followed by four bilinear up-sampling layers. The feature vectors extracted by the encoder are passed directly to the subsequent up-sampling layers in the decoder to infer the depth maps. The loss function is a combination of Structural Similarity (SSIM) loss (Wang et al., 2004) and dense depth loss. The architecture is trained on Stanford2D3D and Matterport3D image sets from the 3D60 dataset (Zioulis et al., 2018).

3.2.3 Materials Recognition

The material estimation is based on network from (Heng et al., 2022), which extracts features from different patch sizes using a self-attention strategy. The network selects the appropriate patch size based on the input image, rather than relying on a fixed patch size for the entire dataset. A set of attention masks (A_1, A_2, A_3, A_4) is predicted and normalized from the final transformer layer to aggregate the features. The merged features are then passed through a feature pyramid network to recover the spatial structure and predict material labels for each pixel. Figure 3.2 shows the output classes of the material recognition module. The network is trained on material database (LMD) (Schwartz and Nishino, 2019).

3.2.4 3D Semantic Scene Completion (SSC)

For 3D semantic scene completion (SSC), we follow the approach of (Kim et al., 2022) and (Dourado et al., 2021). A 3D voxel structure is reconstructed by projecting points from the estimated depth maps into 3D space. To cover the entire 360 surroundings, the 3D coordinate system is partitioned into eight overlapping view parts from the scene's center. Semantic scene completion is then applied to each region using EdgeNet (Dourado et al., 2021), which employs weights pre-trained on the SUNCG dataset (Song et al., 2017) and fine-tuned on the NYUv2 dataset (Silberman et al., 2012). EdgeNet is designed to predict scene semantics and reconstruct occluded parts of the scene. The outputs are merged into a single complete scene using EdgeNet360 (Kim et al., 2022), resulting in a fully reconstructed 3D model with semantic labels. Finally, the material labels predicted in Section 3.2.3 are assigned to the 3D semantic labels within Unity.

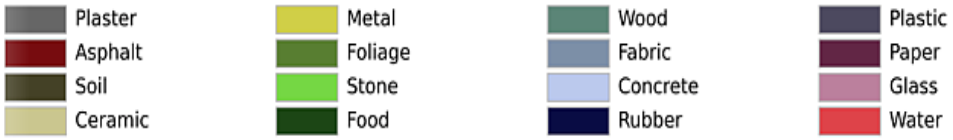

FIGURE 3.2: Materials classes by the proposed material recognition module.

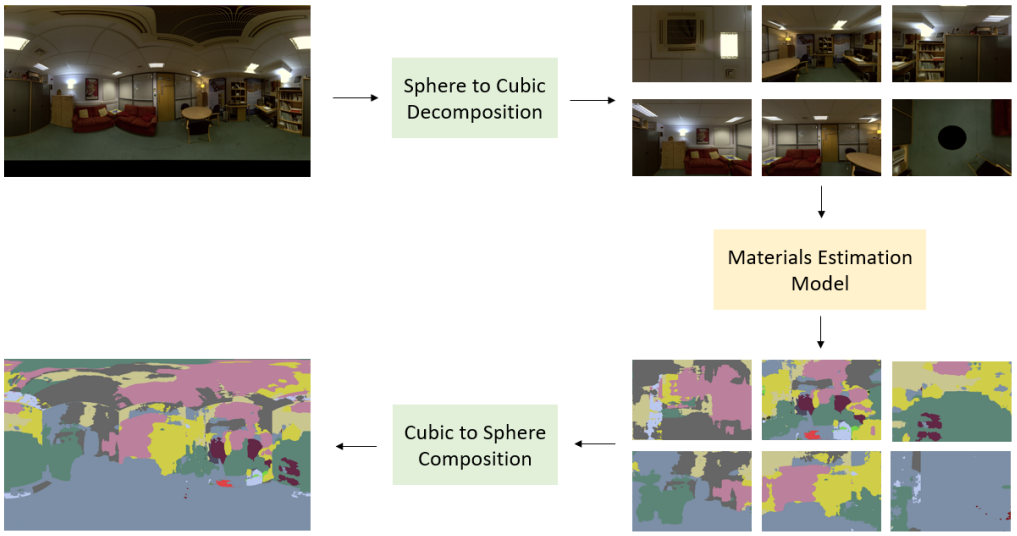

FIGURE 3.3: Sphere to cubic decomposition and composition on MR scene.

3.2.5 Sphere to Cubic Decomposition and Composition

The material estimation process described in Section 3.2.3 only accepts 2D RGB images. Therefore, since the proposed pipeline uses 360° RGB inputs, a cubic projection is applied as a preprocessing step. Each 360° RGB input is decomposed into six segments, with overlapping regions at the boundaries (Kim and Hilton, 2015). These segments are then fed into the material recognition model. After material properties are estimated for each segment, the six cubic parts are recomposed into a complete 360° image. Figure 3.3 illustrates this process.

3.2.6 Sound Rendering in VR Space

The reconstructed 3D semantic scene is imported into Unity [1], where the Steam Audio plug-in Steam Audio[2] is used for room acoustics simulation and sound rendering. A virtual sound source and listener are placed within the scene, with sound rendering including the use of head-related transfer functions (HRTFs). From this configuration, we rendered an exponential sine-sweep method (ESS) mono sound in the virtual rooms

[1] https://unity.com/(accessed in 2024)
[2] https://valvesoftware.github.io/steam-audio/(accessed in 2024)

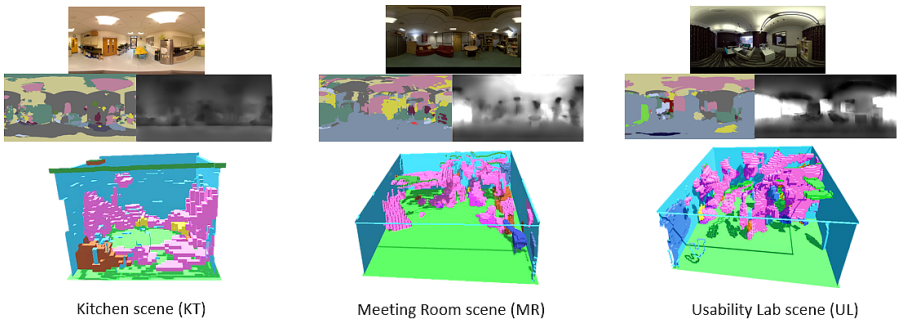

FIGURE 3.4: Visualisations of 3D SSC models constructed from CVSSP dataset samples.

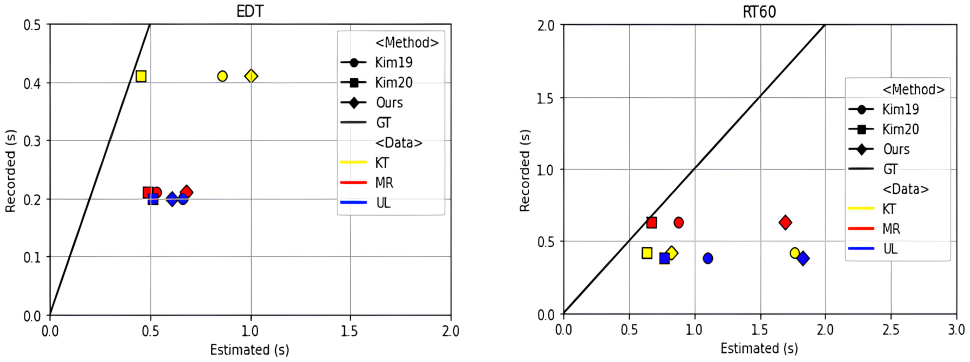

FIGURE 3.5: EDT and RT60 results on CVSSP data related to ground truth.

and we extracted the RIRs following Farina method (Močnik, 2023; Farina, 2000, 2007). The ESS method is advantageous because it varies frequency exponentially over time, making it easier to capture sound propagation across different frequency bands. The RIRs are segmented into direct sound, early reflections, and late reverberations (Yu and Kleijn, 2020). We evaluate early decay time (EDT) and reverberation time (RT60) as objective measures of the virtual environment's acoustic performance. EDT is estimated using the slope of the decay curve, determined from a fit between 0 and -10 dB, and the decay time is calculated from the slope as the time required for a 60 dB decay (Barron, 1995). RT60 measures the time required for the energy to decay by 60 dB, providing insights into the room's size, boundaries, and material absorption properties (Rungta et al., 2016). The average EDT and RT60 values are reported across six octave bands between 250 Hz and 8 kHz.

3.3 Results and Observations

The proposed pipeline is tested on CVSSP [3]. We choose this dataset because it consists of five indoor scenes with 360° images and ground-truth acoustic parameters measurement. This dataset is particularly suitable for assessing both the visual semantics and the acoustic properties in a virtual space. For our preliminary work, three scenes are selected: 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.

In the virtual environments generated for the scenes with 3D models, we analyse the RIRs to assess the spatial sound within the 3D space. Figure 3.4 presents sample outputs from the CVSSP dataset, illustrating the reconstructed 3D models. Upon inspection, we observe that the quality of these models is suboptimal, primarily due to the limitations of the mono depth input. The depth maps failed to consistently capture the correct scale, leading to inaccuracies in the geometry of the reconstructed scenes. Furthermore, we observe that the material estimation network produces errors in its predictions. For example, the sofas in the MR scene are predicted as wood material, as shown in Figure 3.3.

In addition to evaluating the visual quality of the models, we measured the EDT and RT60 values and compared these acoustic parameters with the results from Kim19 (Kim et al., 2019) and Kim20 (Kim et al., 2022), where stereo image pairs are used for the reconstructions, as well as with the ground truth measurements. Figure 3.5 illustrates the results for both EDT and RT60. A comparison of our results with those of (Kim et al., 2019, 2022) revealed noticeable differences, particularly in the accuracy of the acoustic parameters.

The proposed method outperforms Kim19 in EDT for the UL scene and shows better RT60 results compared to Kim19 for the KT scene. However, the RT60 values for the MR and UL scenes are significantly higher than expected. We identified a scale issue in depth estimation for these cases; the actual MR and UL rooms are much smaller with lower ceilings than most rooms in the training datasets. Among the evaluated methods, Kim20 achieved the closest performance to the ground truth (GT).

[3] http://3dkim.com/research/VR/index.html(accessed in 2024)

3.4 Discussion and Summary

In this chapter, we demonstrate the ability to generate an audio-visual VR space from a single 360° RGB input. By integrating various components, including depth and material estimation, and importing the 3D SSC model to a VR gaming platform for sound rendering, we address the first research question (RQ1):

- RQ1: How can we generate an audio-visual VR space from a single 360 ° RGB input?

As outlined in Section 3.2, we proposed a method capable of generating a complete VR space from a single panoramic RGB input. Our approach integrates multiple components essential for constructing an immersive VR environment. These components include mono depth maps for spatial understanding, estimated materials for realistic acoustic rendering, and 3D SSC models for generating virtual room layouts including semantics. The 3D models are imported into the Unity platform, enabling spatial sound rendering and RIR extraction and evaluation. This proposed method simplifies the traditionally resource-intensive process of audio-visual scene generation while providing a scalable solution for various VR applications.

Upon investigating our results, we identified several challenges that affect the VR output. These include limitations with depth input accuracy and the performance of the estimated 3D SSC model. Additionally, discrepancies between estimated materials and those supported by the Steam Audio plug-in complicate object-level acoustic assignments. To address these challenges, we focus on minimizing sources of error and refining the proposed VR space. Specifically, we replace EdgeNet360 with a more advanced model to improve the SSC component. Furthermore, we incorporate stereo depth inputs, which offer greater depth accuracy compared to the previously used mono depth maps. To ensure a fair comparison with state-of-the-art methods, we adopt similar acoustic materials.

We also propose integrating RGB features alongside depth information to enhance 3D SSC model prediction performance, as suggested in SSC literature involving RGB-D perspective inputs. We hypothesise that improving semantic scene completion yields more accurate room geometry estimations, ultimately leading to enhanced predictions of acoustic parameters, particularly EDT and RT60. Therefore, the VR space will be enhanced with both visual and audio cues.

In the following chapters, we investigate several methods to answer research questions RQ 2 and RQ 3:

1. RQ 2: How can the inference capabilities of pre-trained SSC model on perspective images be extended to a single 360° RGB-D input?

2. RQ 3: What is the impact of the generated 3D scene from 360° RGB-D input on acoustic parameters, including early reflections and late reverberations?

To address these questions, in this project, we design 3D SSC deep learning models capable of producing more accurate 3D scene semantics. Chapter 4 and Chapter 5 details the deep learning SSC models, which aim to improve SSC performance using perspective camera inputs. Subsequently, Chapter 6 demonstrates the extension of the model developed in Chapter 5, which processes both RGB and depth features to infer a comprehensive 3D SSC model from 360° RGB-D input, leveraging the CVSSP dataset. Furthermore, we perform an evaluation of RIR acoustic parameters, such as EDT and RT60.

Chapter 4

3D Semantic Scene Completion from a Depth Map with Unsupervised Learning for Semantics Prioritisation

4.1 Motivation and Contributions

In Chapter 3, we illustrated that the 3D model produced by the Semantic Scene Completion (SSC) component significantly impacts scene fidelity. In this chapter, we explore and experiment with constructing SSC model components to contribute to answering the second research question, RQ 2. We begin by incorporating depth-only input from perspective cameras to construct a comprehensive 3D model, including complete scene semantics. SSC aims to build a comprehensive 3D representation from a partial scene view, predicting both volumetric occupancy and object categories (Song et al., 2017). A prime example is the SSCNet (Song et al., 2017), which combines scene completion (SC) and scene semantics (SS). It demonstrates that these two aspects are intrinsically linked and mutually beneficial (Song et al., 2017; Roldao et al., 2022).

SSC differs from 3D reconstruction, which primarily focuses on reconstructing the visible surfaces of a scene. SSC extends beyond this by aiming to understand the semantics of both visible and occluded objects within the scene. As highlighted in (Pan et al., 2023), predicting object semantics within the 3D space is a non-trivial task, given the complexity of inferring information about occluded or partially visible objects. The challenges of prediction within 3D space include dataset imbalances, intraclass diversity, and interclass ambiguity, as illustrated in Figure 1.1. In this chapter, we address

research question RQ 2a by focusing on the data imbalance problem that affects semantic learning in indoor scenes. Our primary objective is to enhance the performance of deep learning models in tackling the SSC problem using single depth input. The proposed approach is designed to generalise across different sensor types that capture depth information.

The data imbalance problem occurring in indoor scenes within 3D space is attributed to the significant representation of empty spaces and the varying quantity of objects at unit level (e.g., voxel level). Although several studies have attempted to address data imbalance between empty and occupied voxels through weighted loss functions (Song et al., 2017; Zhang et al., 2019; Li et al., 2019b; Tang et al., 2022; Dourado et al., 2021, 2022), they often overlook category imbalance within datasets, as discussed in Section 2.3.3. To overcome this limitation, we propose a novel re-weighting method that improves the SSC task. Our method uniquely combines two class re-balancing approaches (re-sampling and class-sensitive learning) and smooths the weights using unsupervised clustering algorithms.

Furthermore, upon reviewing the methodologies employed in training and evaluating SSC models, we observed that most studies, including those by (Wang et al., 2023; Chen et al., 2020; Song et al., 2017; Li et al., 2019b, 2020a; Wang et al., 2022; Liu et al., 2018), utilise the hold-out procedure (further details in Section 2.3.5). This approach typically involves splitting the dataset into two subsets: a training set and a hold-out set, which is reserved for performance evaluation. The machine learning literature offers K-fold cross-validation as a common technique to measure model performance (Nti et al., 2021) and ascertain the model's generalisation error (Anguita et al., 2005; Blum et al., 1999). To the best of our knowledge, no SSC studies in the literature have reported model performance uncertainty by presenting average results alongside standard deviations. Such reporting is vital for establishing reliable and generalisable models performance.

We summarize our contribution as following:

- We investigate different re-weighting methods and introduce a novel re-weighting approach in our loss function to address the imbalance between occupied and empty voxels. Moreover, this method effectively handles imbalances across different categories within the occupied voxels, by leveraging unsupervised learning clustering algorithms.
- We demonstrate the quantification of model performance uncertainty through K-fold cross-validation.
- We compare our approach with state-of-the-art (SOTA) models, and our results significantly surpass SOTA in semantics completion in two public benchmark datasets, with only single depth and similar input and output resolution.

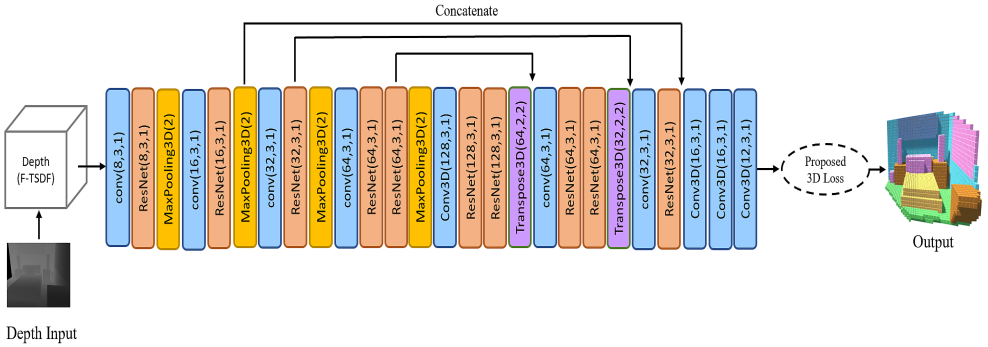

FIGURE 4.1: The architecture design with single encoded depth input (F-TSDF). The network is based on encoder-decoder 3D CNN convolutions and residual modules.

4.2 Methodology

The 3D U-Net CNN implementation in this work employs encoder-decoder blocks, built on (Dourado et al., 2021). Figure 4.1 shows the network architecture. The network includes residual blocks in place of standard convolutions to mitigate the degradation problem often encountered in deep networks. These residual modules incorporate skip connections, enabling the model to correct prediction errors by adding the original input to the residual output. This facilitates the identification of differences between the input and output, enhancing the network's learning efficiency (He et al., 2016a). For volumetric resolution, we adopt the approach used in (Song et al., 2017; Zhang et al., 2018a; Dourado et al., 2021, 2022), employing F-TSDF for geometrical data representation within 3D space, with dimensions of $240\times144\times240$. The model generates an output with a four-dimensional structure sized $60\times36\times60\times12$. The 12 channels represent the dataset classes ranging from 0 to 11. Class 0 represents an empty space, while the remaining classes correspond to different object categories in NYUv2 (Silberman et al., 2012) and NYUCAD (Firman et al., 2016) datasets, such as ceiling, floor, wall, window, chair, bed, sofa, table, TVs, furniture., and object.

4.2.1 Unsupervised Clustering for Class Re-weighting

In a review by (Roldao et al., 2022), the cross-entropy (CE) loss emerges as a preferred loss function for SSC models. Nevertheless, SSC poses unique challenges due to the density of empty data and the existence of occluded portions of objects within scenes, making it difficult to investigate feature relationships in complex settings. While the standard CE loss is commonly used in different contexts, it has well-known drawbacks of treating all the classes equally even when the data are unbalanced. This is because it presumes a uniform significance distribution across all samples and classes (Ho and Wookey, 2019). It measures the affinity between the probability softmax outputs at the end of each forward propagation of the CNN model and the corresponding ground

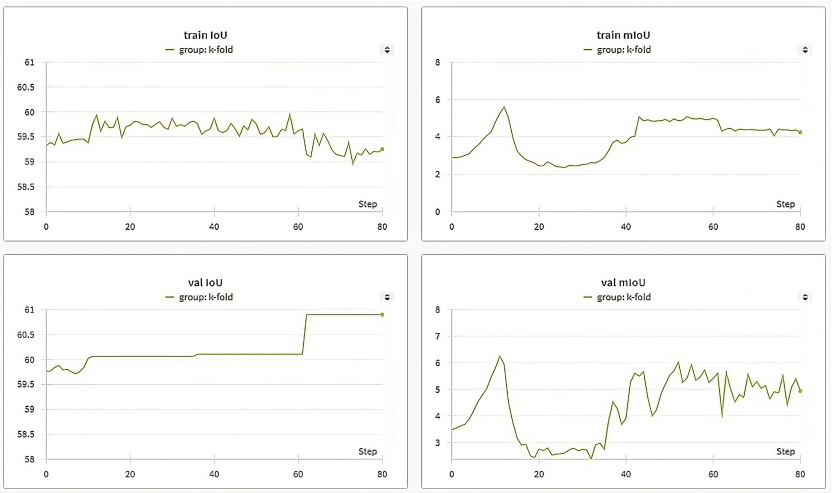

FIGURE 4.2: Training and validation IoU and mIoU curves using the ICF weighting method on the NYUCAD dataset.

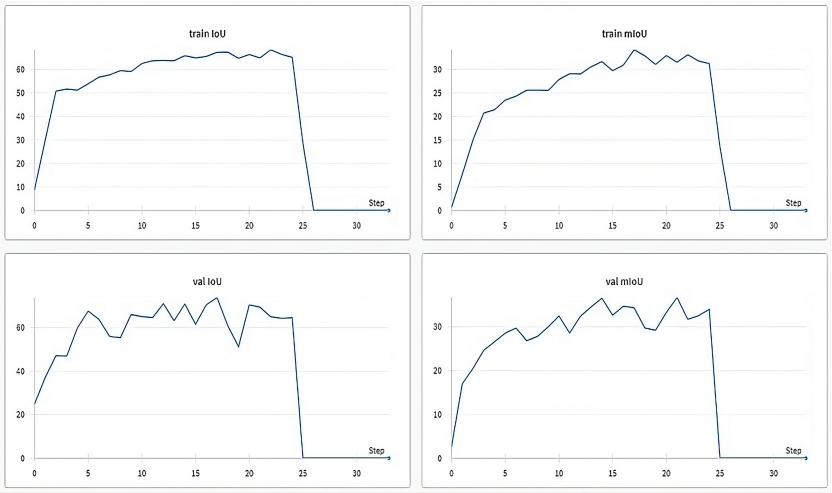

FIGURE 4.3: Training and validation IoU and mIoU curves by employing inverse rank weighting method on NYUCAD dataset. After epoch 25 the measures dropped to zero due to unstable training.

truth (Ho and Wookey, 2019; Song et al., 2016). The aim of the CE loss is to minimize the overall error. However, due to data imbalances, the model weights are frequently adjusted in favor of the majority classes, with infrequent updates for the minor classes. As a result, the misclassifications related to the majority classes have a more pronounced effect on loss minimization which leads to unstable training (Ho and Wookey, 2019). In the context of indoor scenes for SSC problem, data imbalances are common. They often arise from variations in the quantities of data representing empty spaces and other objects within the scenes. The benchmark NYU datasets had the class imbalance problem with significant variations in label distributions with approximately 95% of voxels are unoccupied, and 5% being occupied, as discussed in Section 2.3.1. Common methods used in imbalanced situations are re-sampling and re-weighting (Pan et al.,

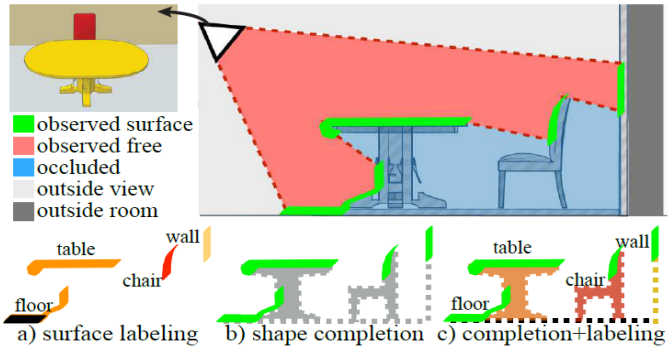

FIGURE 4.4: Regions categories in the scene according to Song et al.(Song et al., 2017) definition .

2023; Zhang et al., 2023). In this work, we evaluate the performance of our model using two class re-balancing methods. Initially, we train the model using the CE with undersampling of occluded empty voxels as proposed by (Song et al., 2017). While this approach enhances the scene completion aspect by balancing between occupied and empty voxels, it treats all scene semantics equally. This uniform treatment led to suboptimal performance for the underrepresented classes. To address this, we employ class-sensitive learning with re-weighting strategy. Specifically, we adopt inverse class frequency (ICF) weights for each voxel as one of the common weighting methods (Eigen and Fergus, 2015; Ronneberger et al., 2015; Li et al., 2019b). However, we observe that applying this method in our research context results in poor performance, characterized by an underfitting problem. The model struggled to learn general patterns within the scenes, resulting in inaccurate scene semantic predictions, as illustrated by the Intersection over union (IoU) and mean Intersection over Union (mIoU) curves in Figure 4.2. Subsequent investigations involve weighting the classes with inverse ranks ranging from 1 to 12 based on their distribution in the dataset. This strategy initially impacts the model's learning positively, but after a few epochs, the loss values diverge to NaN, indicating unstable training caused by the weighting scheme. Consequently, the IoU and mIoU scores dropped to zero, as shown in Figure 4.3. While the ICF method has been successful in other domains, it fails in our context due to the extreme imbalance in our data within complex scenes. This finding aligns with studies on extremely unbalanced datasets, such as those mentioned in (Cui et al., 2019; Sudre et al., 2017). However, our observations from the methods above (re-sampling and inverse ranking weights) inspire us to combine their strengths. By integrating them and utilising unsupervised learning clustering algorithms which perform less aggressive weighting scheme and moderate the weights associated with each class. Therefore, the penalty magnitude is smoothed over the classes (diminished on minor classes and raised on major classes). This strategy lead to a balanced learning approach that emphasizes minor classes without overlooking the major ones. Rather than addressing each class independently, we group them by clustering algorithms based on their label distribution in the dataset, resulting in a more uniform and generalisable representation. In

the representation of indoor scenes, we identify various voxel types, as described in (Song et al., 2017). For a visual reference, see Figure 4.4. The observed surface, represented by the green area in the figure, is denoted by n^s. The occluded occupied and occluded empty types are both represented by the blue area, labeled n^o and n^e, respectively. Empty voxels in the red, gray and light gray areas are labeled as n^f. These voxel types are captured by the set $N = \{n^s, n^o, n^e, n^f\}$, each uniquely associated with a voxel type. In alignment with (Song et al., 2017), we perform undersampling of empty class in occluded regions to balance the learning between occupied and non-occupied voxels in these areas. This is denoted as $n^e \overset{R}{\leftarrow} j$. The empty voxels denoted as n^f, within the red, gray, and light gray areas in Figure 4.4 are ignored in the SSC evaluation as per (Song et al., 2017), and thus, they do not contribute to our loss design. Therefore, our intended subsample set for the scenes is $A = \{n^s, n^o, j\}$. Subsequent to this, we apply the common clustering algorithm, K-means (Hartigan and Wong, 1979), to all voxels in the dataset where each scene have A set representation. This is done using a predetermined number of clusters, denoted as k. To determine the optimal number of clusters, we employ both the elbow method and the Silhouette method (Syakur et al., 2018; learn developers, 2007–2024; Rousseeuw, 1987; Chiang and Mirkin, 2010). Based on the results of these methods, the optimal number of clusters is determined to be $k = 3$, as shown in Figures 4.5 and 4.6. The elbow method involves examining various values of k and identifying the elbow point, which represents the optimal k value where the sum of squared errors (SSE) has significant decrease between the points and cluster's centroid. In contrast, the Silhouette method assesses the difference between within-cluster cohesion and separation from other clusters. Higher Silhouette scores indicate better-defined clusters. These scores range from -1 to 1, a score near zero suggests that an entity could belong to a different cluster, while a score close to -1 indicates possible misclassification. A score close to 1 suggests that the clustering is well-structured and accurate. The Silhouette Coefficient score is defined for each sample and is composed of two components: a the mean distance between a sample and all other points within the same class, and b the mean distance between the sample and all points in the nearest neighboring cluster. The Silhouette Coefficient score is calculated using Equation 4.1:

$$s = \frac{b - a}{\max(a, b)}. \tag{4.1}$$

Therefore, we use the K-means algorithm with three clusters, the data are grouped according to their similar distributions, as shown in Figure 4.7. To further validate these groupings and explore potential insights, we employ an additional algorithm that does not require prior knowledge of the number of clusters. Specifically, we use Density-Based Spatial Clustering of Applications with Noise (DBSCAN) (Ester et al., 1996). The key parameters for the DBSCAN algorithm are epsilon, which defines the maximum distance between two points for them to be considered neighbors, and minimum samples, which represents the minimum number of points required to form a dense region. We optimise the algorithm's configuration by searching for the optimal epsilon and

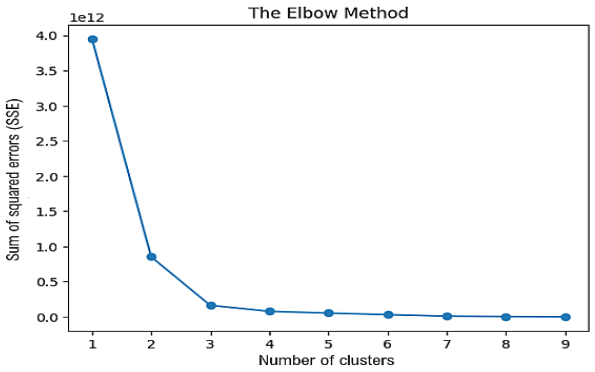

FIGURE 4.5: Illustration of elbow method to select the optimal number of K-means clusters k, where the sum of squared errors between the cluster points and it's centroid is sharply decreased. The X-axis represents different number of clusters k, while the Y-axis denotes the sum of the square error.

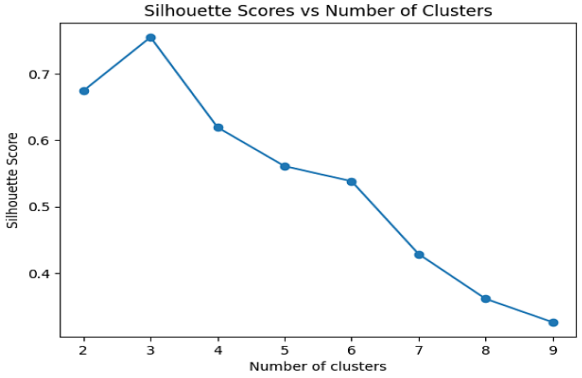

FIGURE 4.6: Illustration of Silhouette method to select the optimal number of K-means clusters k, where segregation between the clusters and the cohesion should be the highest represented by the Silhouette Coefficient score. The X-axis represents different number of clusters k, while the Y-axis denotes Silhouette Coefficient scores.

minimum sample values, identifying 0.4 for epsilon and 1 for the minimum sample size, which achieves the highest Silhouette score of 0.75. This optimization results in three clusters that are equivalent to the results of the K-means algorithm. We denote the clusters as $C = \{C_1, C_2, \ldots, C_k\}$ where k is the number of clusters. For each cluster C_k, we compute the average distribution μ as denoted in Equation 4.2:

$$D(C_k) = \mu C_k \tag{4.2}$$

We assign weights for the clusters, w, based on the inverse order of their average distribution. Specifically, clusters with a lower average distribution receive higher weights, and vice versa. Consequently, the voxel classes are categorized under k distinct weights, represented by $w = (w_1, w_2, \ldots, w_k)$. For each voxel v belonging to cluster C_k, the weight assigned to it, w_v is equivalent to the weight of the cluster, w_k. This relationship is captured in the Equation 4.3:

$$\forall v \in C_k, \quad w_v = w_k \tag{4.3}$$

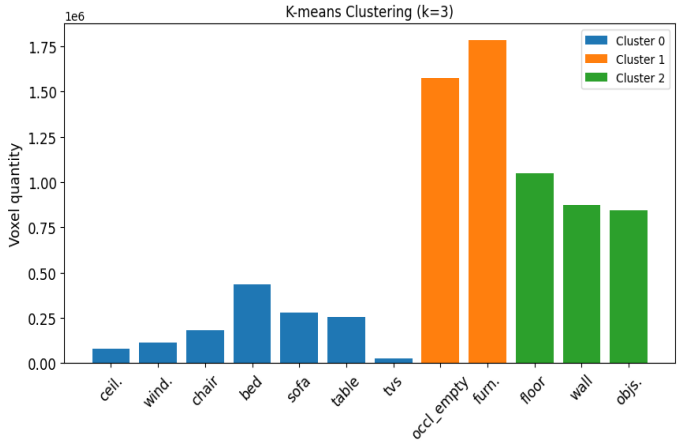

FIGURE 4.7: Clustering of NYU voxels labels using K-means Algorithm. The X-axis represents the various classes, while the Y-axis denotes the quantity of voxels.

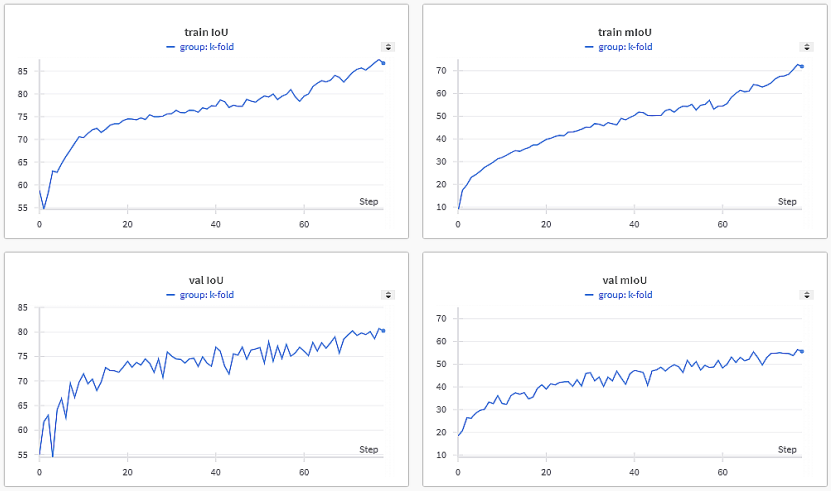

FIGURE 4.8: Training and validation IoU and mIoU curves using the proposed method on NYUCAD dataset.

In our approach, we compute the CE loss function between the predicted class label p and the actual class label y for the voxels within a scene. Let each voxel v be an element in the set A, the predicted class label and actual class label for voxel v are denoted as p_v and y_v respectively. Each voxel label is given a weight w_v, as defined in Equation 4.3. The weighted CE loss function is outlined in Equation 5.6:

$$L(p,y) = -\sum_{v=1}^{A} w_v \,.\, y_v \,.\, log p_v \tag{4.4}$$

This re-weighting method moderates the penalty applied to each voxel's label. Consequently, this results in a more stable training process and a more generalised representation. Figure 4.8 shows the IoU and mIoU curves.

4.2.2 Uncertainty Quantification of Model Performance Using K-fold Cross Validation

In this research, we employ K-fold cross-validation during the training process rather than relying on fixed splits and averaging results across multiple runs. This approach divides the dataset into k_f folds of equal size. Each fold is sequentially used as validation data, while the remaining $k_f - 1$ folds are used for training. The process is repeated k_f times, and the performance results from all folds are averaged to provide a single performance estimate for the model (Stone, 1974; Wong, 2015; Rodriguez et al., 2009). This method addresses a gap in SSC model assessment by eliminating biases associated with selecting single-score performance that representing the best results such in (Zhang et al., 2018a). It also enables the monitoring of model fitting behavior through the use of a validation set. Reporting average results with standard deviations is a crucial step toward building reliable and generalised models. Moreover, the minimal variation in reported scores across different models highlights the difficulty of effectively comparing performance without averaging the scores over multiple runs and providing the standard deviations.

However, given the unbalanced data distribution within complex scenes in the NYU datasets, choosing an ideal value for k_f proves challenging. To address this issue, we use k_f=3 due to training time constraints and the additional complexity introduced by the unbalanced data distribution. We then evaluate the testing set against each fold, average the results, and report the standard deviations for Precision, Recall, and IoU for the SC task, as well as the mIoU for the SS task as shown in Table 4.1 and Table 4.2.

4.3 Experiments

4.3.1 Implementation Details

The implementation of this work is divided into three main phases: preprocessing, training and validation, and evaluation. The code can be accessed at: `https://github.com/MonaIA1/SSC/tree/master`. In this work, we name our model DBNet, which stands for Data Balancing Network due to the use of our proposed loss function.

4.3.1.1 Datasets and Preprocessing

To conduct the experiments, we utilise the benchmark datasets NYUv2 (Silberman et al., 2012) and NYUCAD (Firman et al., 2016). NYUv2 comprises 1,449 realistic RGB-D indoor scenes captured with a Kinect sensor at a resolution of 640×480. The dataset

TABLE 4.1: Results on NYUv2 dataset: our results are averaged with std scores over Prec., Recall, IoU, and mIoU. In the input column, 'D' means depth map.

Method	Input	Res.	Scene Completion (SC)			Semantic Scene Completion (SS)										
			Prec.	Recall	IoU	ceil.	floor	wall	win.	chair	bed	sofa	table	tvs	furn.	objs	mIoU
FFNet(Wang et al., 2022)	RGB-D	(60,60)	89.3	78.5	71.8	44.0	93.7	41.5	29.3	36.2	59.0	51.1	28.9	26.5	45.0	32.6	44.4
3D Sketch(Chen et al., 2020)	RGB-D	(60,60)	85.0	81.6	71.3	43.1	93.6	40.5	24.3	30.0	57.1	49.3	29.2	14.3	42.5	28.6	41.1
TS3D (Garbade et al., 2019)	RGB-D	(240,60)	-	-	60.0	9.7	93.4	25.5	21.0	17.4	55.9	49.2	17.0	27.5	39.4	19.3	34.1
EdgeNet (Dourado et al., 2021)	RGB-D	(240,60)	76.0	68.3	56.1	17.9	94.0	27.8	2.1	9.5	51.8	44.3	9.4	3.6	32.5	12.7	27.8
AICNet (Li et al., 2020a)	RGB-D	(240,60)	62.4	91.8	59.2	23.2	90.8	32.3	14.8	18.2	51.1	44.8	15.2	22.4	38.3	15.7	33.3
CCPNet (Zhang et al., 2019)	D	(240,240)	74.2	90.8	63.5	23.5	96.3	35.7	20.2	25.8	61.4	56.1	18.1	28.1	37.8	20.1	38.5
VVNet (Guo and Tong, 2018)	D	(240,60)	69.8	83.1	61.1	19.3	94.8	28.0	12.2	19.6	57.0	50.5	17.6	11.9	35.6	15.3	32.9
PALNet (Li et al., 2019b)	D	(240,60)	68.7	85.0	61.3	23.5	92.0	33.0	11.6	20.1	53.9	48.1	16.2	24.2	37.8	14.7	34.1
ESSCNet (Zhang et al., 2018a)	D	(240,60)	71.9	71.9	56.2	17.5	75.4	25.8	6.7	15.3	53.8	42.4	11.2	0.0	33.4	11.8	26.7
SSCNet (Song et al., 2017)	D	(240,60)	57.0	94.5	55.1	15.1	94.7	24.4	0.0	12.6	32.1	35.0	13.0	7.8	27.1	10.1	24.7
Baseline(Ours)	D	(240,60)	84.0±1.42	76.6 ±1.72	66.6±0.71	30.6	92.9	46.6	0.0	9.6	55.5	39.6	12.8	0.0	39.3	17.7	31.3±3.73
DBNet(Ours)	D	(240,60)	79.3±0.99	83.3 ±0.84	68.1±0.51	48.9	92.8	49.2	0.0	31.7	61.4	56.1	29.2	0.0	33.9	19.3	38.4±0.17

TABLE 4.2: Results on NYUCAD dataset: our results are averaged with std scores over Prec., Recall, IoU, and mIoU. In the input column, 'D' means depth map.

Method	Input	Res.	Scene Completion (SC)			Semantic Scene Completion (SS)											
			Prec.	Recall	IoU	ceil.	floor	wall	win.	chair	bed	sofa	table	tvs	furn.	objs	mIoU
FFNet(Wang et al., 2022)	RGB-D	(60,60)	94.8	90.3	85.5	62.7	94.9	67.9	35.2	52.0	74.8	69.9	47.9	27.9	62.7	35.1	57.4
3D Sketch(Chen et al., 2020)	RGB-D	(60,60)	90.6	92.2	84.2	59.7	94.3	64.3	32.6	51.7	72.0	68.7	45.9	19.0	60.5	38.5	55.2
TS3D (Garbade et al., 2019)	RGB-D	(240,60)	-	-	76.1	25.9	93.8	48.9	33.4	31.2	66.1	56.4	31.6	38.5	51.4	30.8	46.2
AICNet (Li et al., 2020a)	RGB-D	(240,60)	88.2	90.3	80.5	53.0	91.2	57.2	20.2	44.6	58.4	56.2	36.2	9.7	47.1	30.4	45.8
CCPNet (Zhang et al., 2019)	D	(240,240)	91.3	92.6	82.4	56.2	94.6	58.7	35.1	44.8	68.6	65.3	37.6	35.5	53.1	35.2	53.2
VVNet (Guo and Tong, 2018)	D	(240,60)	86.4	92.0	80.3	-	-	-	-	-	-	-	-	-	-	-	-
SSC-cGAN-GL (Chen et al., 2019)	D	(240,60)	80.7	91.1	74.8	-	-	-	-	-	-	-	-	-	-	-	42.0
PALNet (Li et al., 2019b)	D	(240,60)	87.2	91.7	80.8	54.8	92.8	60.3	15.3	43.1	60.7	59.9	37.6	8.1	48.6	31.7	46.6
SSCNet (Song et al., 2017)	D	(240,60)	75.0	92.3	70.3	-	-	-	-	-	-	-	-	-	-	-	-
Baseline(Ours)	D	(240,60)	88.8±0.70	88.1 ±0.48	79.3±0.56	55.1	93.0	59.9	1.4	41.9	53.4	52.5	31.5	0.0	51.7	24.7	42.3±3.18
DBNet(Ours)	D	(240,60)	86.5±0.91	91.1 ±1.08	79.6±0.13	66.7	93.6	60.7	15.7	51.4	68.9	68.7	45.6	0.0	44.9	29.3	49.6±1.22

is divided into two splits: 795 instances for training and 654 instances for testing. 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 use the high-quality NYUCAD synthetic dataset, which projects depth maps from ground truth annotations and avoids misalignment.

Furthermore, we encode all 2D depth maps from the NYUv2 and NYUCAD datasets into 3D space using F-TSDF (Song et al., 2017). The processed data is saved and can be used across multiple designs. In line with previous works (Song et al., 2017; Liu et al., 2018; Dourado et al., 2022), we align the 3D scenes layout with the Manhattan world assumption, which is related to the direction of gravity. The defined 3D space dimensions are 4.8 meters in width, 2.88 meters in height, and 4.8 meters in depth. With a voxel grid size of 0.02 meters, this configuration results in a volumetric resolution of $240 \times 144 \times 240$ voxels. The TSDF truncation value is set to 0.24 meters, optimizing the balance between detail capture and computational efficiency. We utilise the parallel computing capabilities on Graphics Processing Unit (GPU). The GPU facilitates parallel processing for the F-TSDF volumetric representation and distance calculations between the 3D points within the 3D volumes. Also, the re-sampling of the occluded empty voxels is provided at this stage.

4.3.1.2 Training and Validation

We train from scratch using both NYUv2 and NYUCAD datasets. Our experiments are implemented in the PyTorch framework, utilising one Tesla v100 GPU.For training, we adopt the mini-batch Stochastic Gradient Descent (SGD) with a momentum of 0.9, using a batch size of 4 for training and 2 for validation, and a weight decay of 5×10^{-4}. The OneCycleLR learning rate scheduler is employed, with an initial learning rate set at 0.01. We train the SSC model for 100 epochs, incorporating an early stopping as a regularization method (Moradi et al., 2020) with patience of 15 epochs to prevent overfitting on the training data. Additionally, we implement K-fold cross-validation by randomly dividing the training sets into three folds and save the weights of each trained fold for later evaluation on the testing set.

4.3.1.3 Evaluation Metrics

We adopt Precision, Recall, and IoU as the evaluation measures for the SSC, following the approach of Song et al. (Song et al., 2017). For the SS task, both the observed surface and occluded regions are evaluated. We present the mIoU scores for semantic classes, excluding the empty class. In the SC task, all non-empty voxels are classified as '1', while empty voxels are labeled as '0'. The binary IoU is computed for the occluded

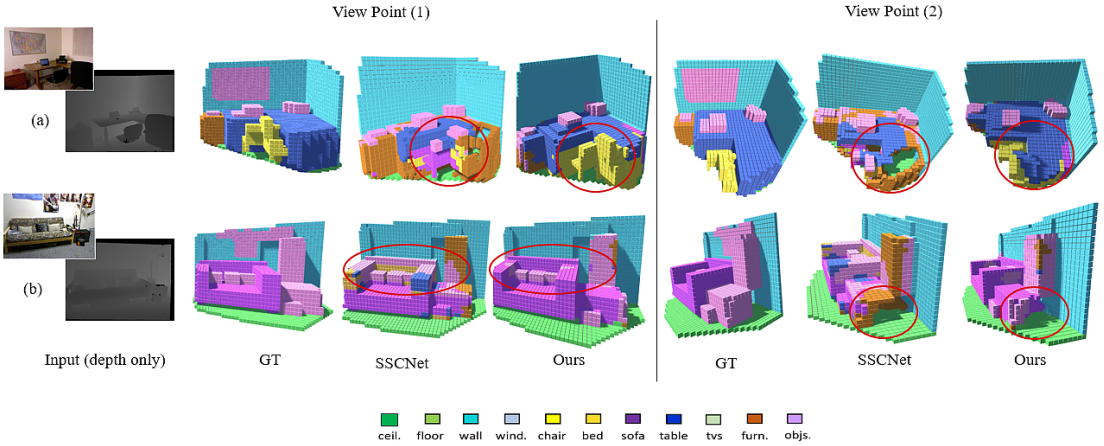

FIGURE 4.9: SSC results from depth maps by NYUv2 dataset. The 3D models displayed from different viewpoints to illustrate the scene completion in the occluded regions and scene semantics. Each object represented by a unique color and circles show the main differences between GT and the predictions by SSCNet model and ours.

regions in the view frustum along with precision and recall measures. We have observed that there's no standardized method for selecting the SC area, leading to slight variations among researchers in the field as we mentioned in Section 2.3.4. 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., 2019b), DDRNet (Li et al., 2019a), and AICNet (Li et al., 2020a), 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 the occluded occupied voxels and re-sampled empty occluded ones. However, mIoU according to (Liu et al., 2024; Li et al., 2020a), is considered more critical than IoU. Finally, the results for all measures are averaged across the K-fold cross-validation to derive the final scores.

4.3.2 Comparisons with State-of-the-Art Approaches

We conduct experiments to evaluate the effectiveness of our proposed method on both SC and SS tasks using the NYUv2 and NYUCAD datasets. In Table 4.1 and Table 4.2, we provide a quantitative comparison of our re-weighting method where weights are smoothed to three values using K-means with $k = 3$, and comparing the results against SOTA approaches. While previous studies report results without specifying the uncertainty in performance, we average our scores across the K folds to better capture and represent the generalisation performance.

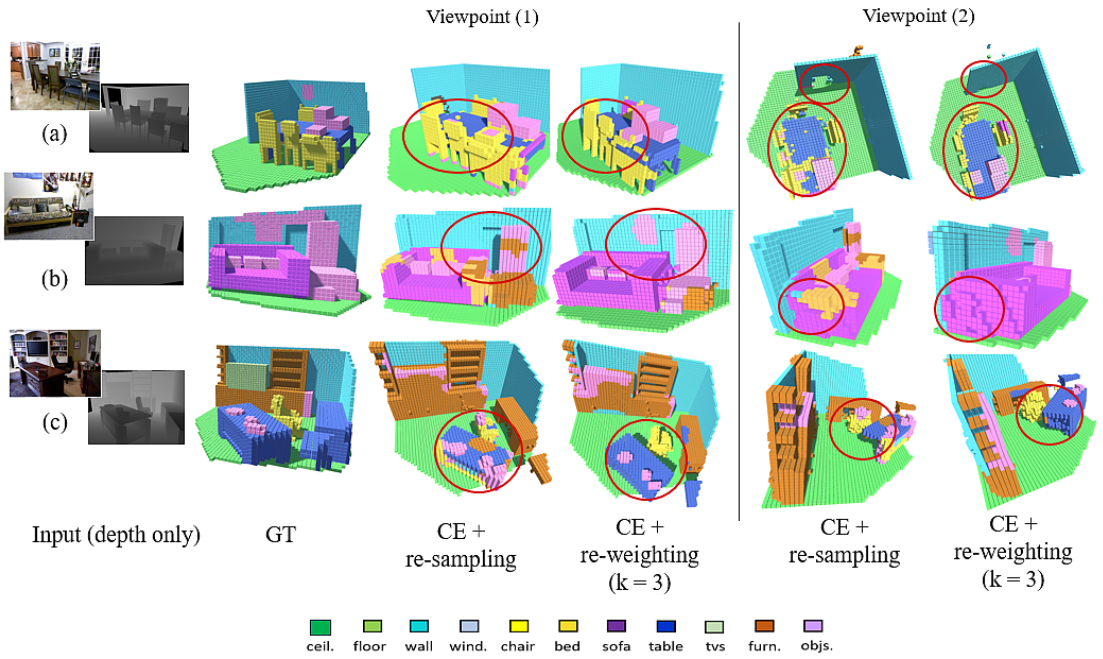

FIGURE 4.10: A visualisation of SSC results with different loss components on NYU-CAD dataset from different viewpoints. From left to right: (1) Input depth; (2) Ground truth (GT); (3) using CE loss with re-sampling method; (4) using CE loss with the proposed re-weighting using K-mean clustering with k=3. Each object represented by a unique color and circles show the main differences between GT and the predictions.

NYUv2 Dataset. Our method significantly outperforms other existing models by a substantial margin on NUYv2 dataset, as shown in Table 4.1. We restrict our comparison to models that use only a single depth input, such as those in (Song et al., 2017; Zhang et al., 2018a, 2019; Li et al., 2019b; Guo and Tong, 2018). Specifically, our approach surpasses (Song et al., 2017; Zhang et al., 2018a; Li et al., 2019b; Guo and Tong, 2018) with considerable gains ranging from 6.8 percentage points (pp) to 13 pp for IoU scores and from 4.3 pp to 13.7 pp for mIoU scores. Furthermore, our loss design proves to be effective, and our model significantly outperforms models with multiple inputs such as (Li et al., 2020a; Garbade et al., 2019), and shows competitive performance to (Chen et al., 2020; Wang et al., 2022).

NYUCAD Dataset In Table 4.2, we also conduct tests on the NYUCAD dataset to validate the generalisability of our proposed method. In the SC task, our method significantly surpasses (Song et al., 2017) and (Chen et al., 2019) with gains of 9.3 pp and 4.8 pp, respectively. In the SS task, our method significantly outperforms (Li et al., 2019b), (Chen et al., 2019), (Li et al., 2020a), and (Garbade et al., 2019) with gains of 3 pp, 7.6 pp, 3.8 pp, and 3.4 pp, respectively. Moreover, our results show competitive performance compared to models with larger resolutions, such as in (Zhang et al., 2019), and to models that use multiple inputs (Chen et al., 2020; Wang et al., 2022).

TABLE 4.3: Ablation studies on loss components performed on NYUCAD dataset.

Method	SC IoU%	SS mIoU%
CE + re-sampling	79.3 ±0.56	42.3 ±3.18
CE + re-weighting ($k = 3$)	79.6 ±0.13	49.6 ±1.22

Additionally, in both the NYUv2 and NYUCAD datasets, our precision and recall scores demonstrate our model's capability to distinguish between occupied and empty voxels. In contrast, the results reported on NYUv2 in (Zhang et al., 2018a) show equal values for precision and recall. This indicates that their model produced an equal number of false positive and false negative predictions, which is not the case with our model.

4.3.3 Ablation Study

To assess the impact of our method's key components, we conduct ablation studies on both NYUCAD dataset. We initiate this assessment by training our model with re-sampling and cross-entropy loss (Song et al., 2017) to balance between occupied and empty voxels. Subsequently, we apply our re-weighting method, proposing three weights based on an unsupervised clustering algorithm. This is illustrated in Table 4.3. We observe an improvement in the average IoU score with 0.3 pp. Furthermore, there is a substantial boost in the mIoU score by 7.3 pp. This clearly demonstrates the effectiveness of our re-weighting method, which targets the differentiation of learning across dataset categories based on their label distribution.

4.3.4 Qualitative Results Analysis

To demonstrate the effectiveness of our re-weighting method in the loss design and its capability to produce more accurate predictions, we present several visualisations from the NYUv2 dataset in Figure 4.9, comparing our approach to the baseline SSCNet (Song et al., 2017) trained with our specified method in Section 4.3.1. Our employed re-weighting strategy yields an improvement in SC, better interclass distinction, and enhanced intraclass consistency, as evidenced in (a) and (b) in Figure 4.9. On the other hand, in Figure 4.10, the predicted 3D models from the NYUCAD dataset are presented, comparing scenarios where our proposed weighting method is applied versus when they are not. Upon implementing the proposed re-weighting method, noticeable improvements in predictions for both the SC and SS tasks are evident. Our DBNet model also demonstrates its capability in completing occluded sections such as wall behind the dining table, and the occluded parts of the chairs in Figure 4.10 (a). From Figure 4.10 (b), we can see the model ability to predict the sofa and objects on it and on the wall. Furthermore, there is a significant boost in intraclass consistency, evidenced by

the improved differentiation between different chairs with different shapes and different tables with different shapes in (a) and (c) in Figure 4.10. However, within the same figure, (c) reveals that the model occasionally confuses the TV class with objects and furniture classes. This confusion arises because TVs share similar features with these classes and represent the least common category in the dataset, making them difficult to recognize from depth information alone. Nonetheless, by visualizing scenes from different views, it becomes evident that our method offers superior shape and semantics completion compared to using re-sampling alone.

4.4 Discussion and Summary

In this chapter, we address the problem of SSC, which involves the joint inference of volumetric occupancy and object categories from a single depth input, providing only a partial view of the scene. Our proposed DBNet SSC model utilises a single depth input encoded with F-TSDF for geometry representation to predict full 3D scenes. This approach enhances the model's adaptability across a variety of depth-sensing devices.

We contribute to overcoming a key challenge in this domain, primarily the inherent imbalance in 3D spatial distributions commonly observed in indoor scenes. To address this issue, we introduce a re-weighting method integrated into the loss function, leveraging the K-means clustering algorithm.

We investigate the implication of various voxel prioritisation strategies by testing different weighting schemes for voxels within scenes from the NYU datasets, as detailed in Section 4.2.1. This investigation guided the systematic development of the proposed re-weighting method, designed to improve the model's ability to learn meaningful semantic representations while mitigating class imbalance issues.

The voxel re-weighting strategy is integrated into the DBNet model to evaluate its impact on SSC model learning. As demonstrated by the qualitative results in Figure 4.10, the model exhibit improved semantic scent completion compared to the baseline model where class prioritisation is absent. Furthermore, the quantitative performance evaluation of the NYUv2 and NYUCAD benchmark datasets, presented in Table 4.1 and Table 4.2, show notable improvements over baseline and comparable state-of-the-art methods.

Our proposed method effectively addresses research question RQ 2a:

- RQ 2a: What is the impact of prioritising voxel weights within the scene on the SSC model learning with depth only input?

FIGURE 4.11: IoU performance on NYUv2 dataset classes using the baseline model and DBNet model with depth-only input.

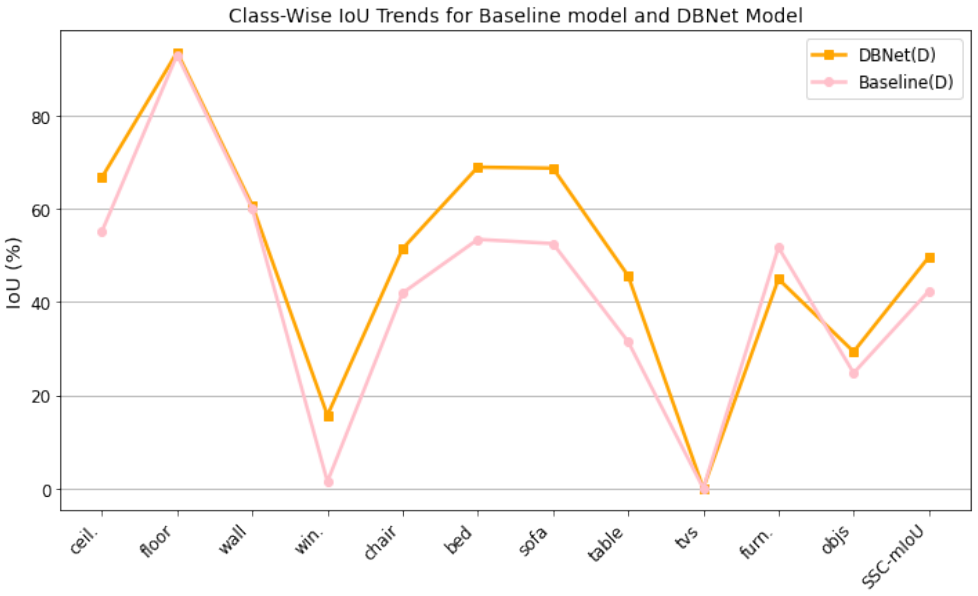

FIGURE 4.12: IoU performance on NYUCAD dataset classes using the baseline model and DBNet model with depth-only input.

These findings demonstrate that voxel prioritisation plays a critical role in enhancing SSC model learning by addressing dataset imbalances, and ultimately boosting the overall semantic scene completion performance.

Despite the advancements in overall IoU and mIoU scores shown in Table 4.1 and Table 4.2, we analyse the impact of our method on scene semantics by visualizing IoU trends on both the NYUv2 and NYUCAD datasets at the class level, as shown in Figure 4.11 and Figure 4.12. Notably, the model demonstrates improvements in classes such as

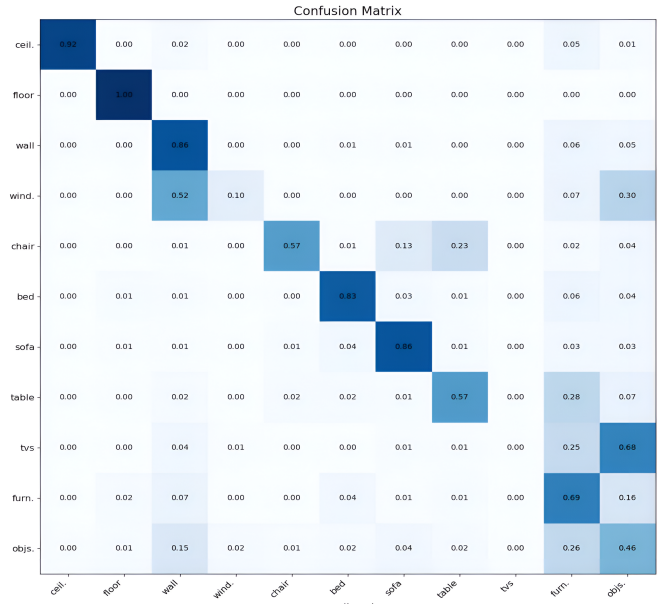

FIGURE 4.13: Confusion matrix of our method on the NYUCAD testing set. The diagonal represents correct predictions, while off-diagonal entries indicate misclassifications. The x-axis corresponds to predicted classes, and the y-axis to ground truth classes.

chair, bed, sofa, table, objects, and ceiling. However, challenges remain in recognizing certain objects within the scene, such as TVs and windows.

Windows are particularly difficult to detect due to their reflective or transparent surfaces, while TVs share visual characteristics with other objects. For example, the model occasionally misclassifies TVs as furniture or other object classes, as qualitatively illustrated in Figure 4.10(c), making them challenging to identify using depth information alone. Additionally, Figure 4.13 illustrates the confusion matrix of our proposed method and highlights how classes are misclassified, as shown in off-diagonal values, demonstrating that TVs are highly misclassified as objects and furniture, while windows are misclassified as walls and objects. Furthermore, in Figure 4.11 and Figure 4.12, we observe a reduction in performance for the furniture class, which is the most dominantly occupied class in the dataset. This decline could be a side effect of the re-weighting method, similar to the observation in other research context (Sinha et al., 2020).

In the next chapter, we address these limitations within the current DBNet design, aiming to enhance the model's performance and provide more accurate SSC predictions.

Chapter 5

Semantic Scene Completion with Multi-Feature Data Balancing Network (MDBNet)

5.1 Motivation and Contributions

Revisiting the challenges in the Semantic Scene Completion (SSC) domain as illustrated in Chapter 1, the partial-view nature of input data complicates the assignment of accurate semantic labels within 3D spaces. This complexity is further exacerbated by factors such as dataset imbalance, intraclass diversity, and interclass ambiguity (Pan et al., 2023). Our DBNet model in the previous Chapter 4 contributes to addressing the class imbalance by introducing a weighted cross-entropy function combined with a re-weighting method based on re-sampling and unsupervised clustering. Although this approach improved the overall mIoU score and the recognition of underrepresented classes such as chairs and tables, DBNet struggled with challenging objects, such as TVs and windows. Windows often feature reflective or transparent surfaces, while TVs share visual characteristics with other categories, such as generic objects, making them difficult to distinguish using depth information alone in datasets with complex scenes like NYUv2 (Silberman et al., 2012) and NYUCAD (Firman et al., 2016). To address these challenges, we incorporate RGB inputs to provide more features to support learning within the network.

In this chapter, we extend the DBNet in Chapter 4 by proposing a dual-head network, utilising transfer learning with projection of RGB semantic features to the 3D space with a combined loss function. We explore various strategies for fusing RGB semantic features. The optimal approach is selected based on average performance scores. Furthermore, inspired by (Zhang et al., 2019; Park et al., 2019; Weder et al., 2020) our approach incorporates the 3D Identity Transformed within full pre-activation Residual

Module (ITRM), an innovative adaptation in the 3D CNN branch of the proposed MDB-Net. This design introduces hyperbolic tangent activation (Tanh) on identity features, enabling effective processing of both positive and negative signals from F-TSDF inputs while normalizing feature distributions between -1 and 1. Furthermore, we answer the research question RQ 2b, by constructing our SSC model and laying the foundation SSC with RGB-D 360°.

We summarise our contributions as follows:

- We propose a hybrid architecture with dual heads to simultaneously learn from multiple data representations of a single scene, leveraging a combined loss function with a re-weighting method proposed in Chapter 4. This design improves learning within the complex scenarios of indoor scenes by incorporating loss from 2D RGB semantics and 3D SSC.
- We evaluate different RGB semantics fusion strategies by assessing the average scores using K-fold cross-validation. This comprehensive analysis facilitates the selection of fusion methods that effectively validate the model's generalisation across diverse scenarios.
- We enhance the overall results by implementing the ITRM block, which includes a hyperbolic tangent activation function applied to identity features. This approach emphasizes positive signals for visible spaces and negative signals for occluded regions and ensures compatibility with the characteristics of the F-TSDF data.
- We compare our method with relevant state-of-the-art (SOTA) models, and our results outperform SOTA in semantic completion on two public benchmark datasets.

5.2 Method

5.2.1 Overall Framework

The architecture of the proposed MDBNet is depicted in Figure 5.1. 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 preprocesses as the form of F-TSDF for data representation within 3D space, which captures geometric information with dimensions of 240×144×240. We leverage the Segformer, a pre-trained transformer model for image semantic segmentation, to extract the 2D semantic features, which are subsequently projected into 3D space. For the 3D input, we adopt the foundational structure of the 3D U-Net CNN, as utilised

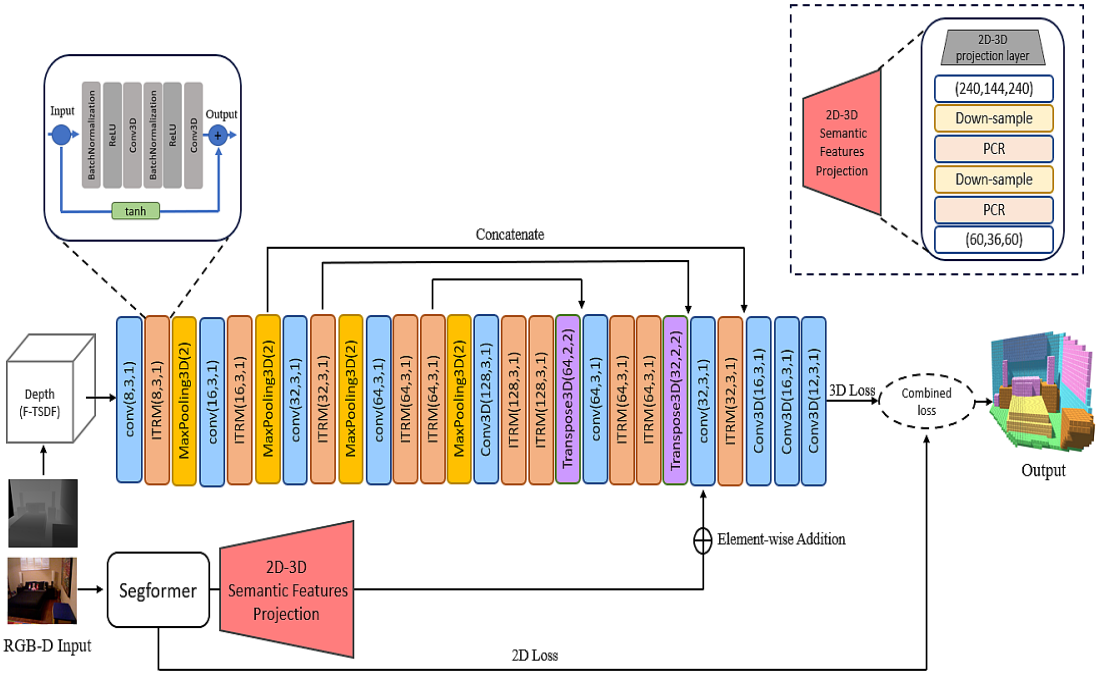

FIGURE 5.1: MDBNet: a multi-feature network with dual heads for processing both 2D RGB semantics and geometric data. The first branch from the bottom utilises a pre-trained Segformer for 2D RGB semantics, incorporating a 2D-3D projection module with nested PCR blocks. The second branch processes geometric data represented by F-TSDF in 3D space, using a 3D CNN that includes an encoder, decoder with ITRM blocks. The network optimises a combined loss, which is a weighted sum of 3D loss and 2D semantics loss.

in (Dourado et al., 2021), with a custom adaptation of the residual block. This adaptation includes adding Tanh on identity features. The model generates an output with a four-dimensional structure sized $60\times36\times60\times12$. 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.

5.2.2 2D Semantic Features

The incorporation of 2D RGB semantic features beside the F-TSDF features, can provide more guidance for SSC model learning. Specifically, RGB semantics add surface features to the objects in scenes, features that are absent in methods relying solely on depth maps as input. Transfer learning emerges as the most effective strategy for this adaptation process. It facilitates the efficient extraction of these RGB semantic features, enabling the system to benefit from learning more diverse features across larger dataset. Consequently, to optimize 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 initializing the decoder's weights with those pre-trained on ADE20K, followed by fine-tuning on the NYU datasets (Wang et al., 2023).

5.2.3 2D-3D Features Projection

Features extracted from 2D RGB images are projected and mapped onto the corresponding coordinates in 3D space by taking the 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 I_{depth}, along with the intrinsic camera matrix $K \in \mathbb{R}^{3\times3}$ and the extrinsic camera matrix $[R|t] \in \mathbb{R}^{3\times4}$ to project a pixel $p_{u,v}$ represented in homogeneous coordinates as $[u, v, 1]^T$ from the 2D image plane to a 3D point $p_{x,y,z}$, also in homogeneous coordinates $[X, Y, Z, 1]^T$. This projection is accomplished using the camera projection equation referenced as Equation 5.1:

$$p_{u,v} = K[R|t]p_{x,y,z}, \tag{5.1}$$

to map the 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 according to the fusion strategies defined below.

5.2.4 3D Features Fusion Strategies

Different fusion methods based on element-wise addition are implemented to assess the model's performance, including early, mid, and late fusions as shown in Figure 5.2, Figure 5.3, and Figure 5.4. 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. Early fusion involve combining the full-resolution projected 3D surface features 240×144×240 with the F-TSDF input prior to their introduction into the 3D network branch. For mid and late fusions, the projected 3D surface features downsampled to align with the resolutions of the network's intermediate 15×9×15 and later 60×36×60 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., 2019a),

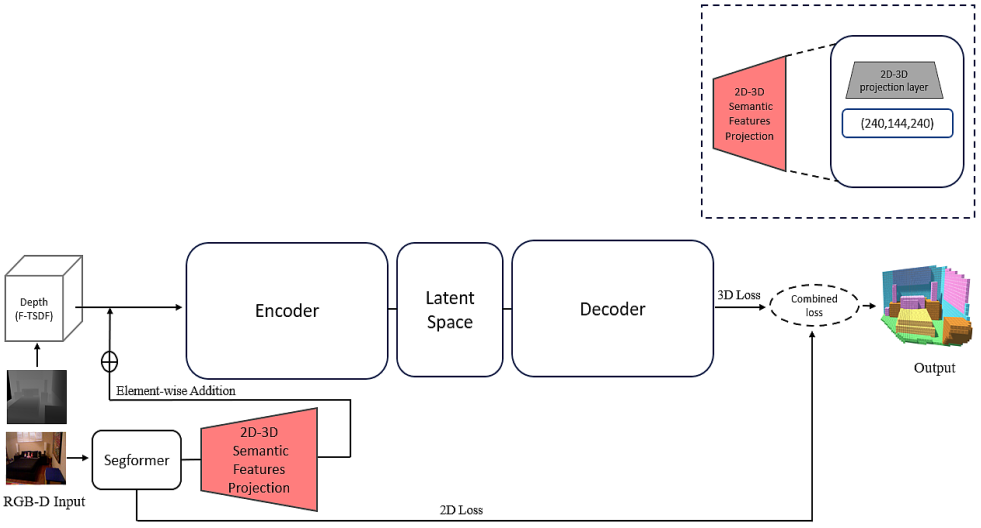

FIGURE 5.2: Early fusion of RGB semantics features in the network.

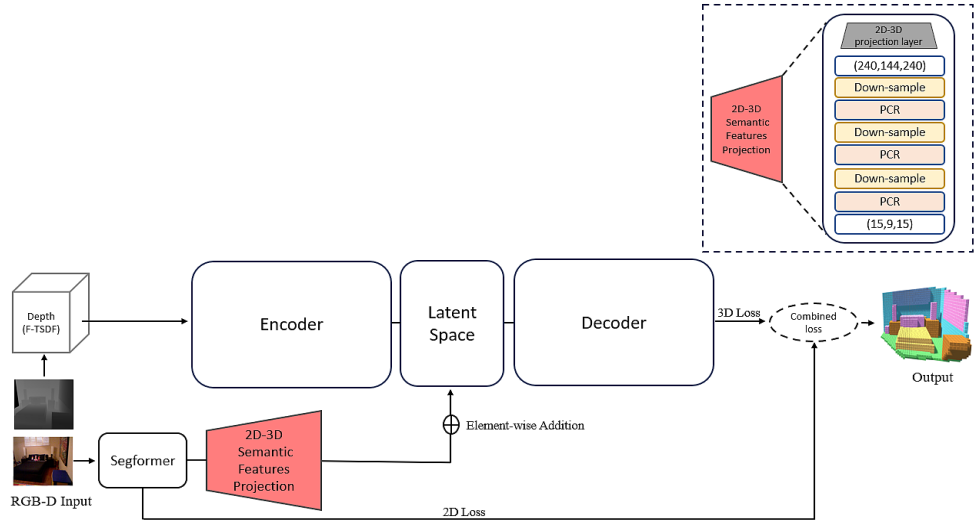

FIGURE 5.3: Middle fusion of RGB semantics features in the network.

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 parameter count relative to standard residual blocks.

5.2.5 Identity Transformed within full pre-activation Residual Module (ITRM)

In this research, we propose a modification to the residual blocks by incorporating a hyperbolic tangent (Tanh) function on the identity features. 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

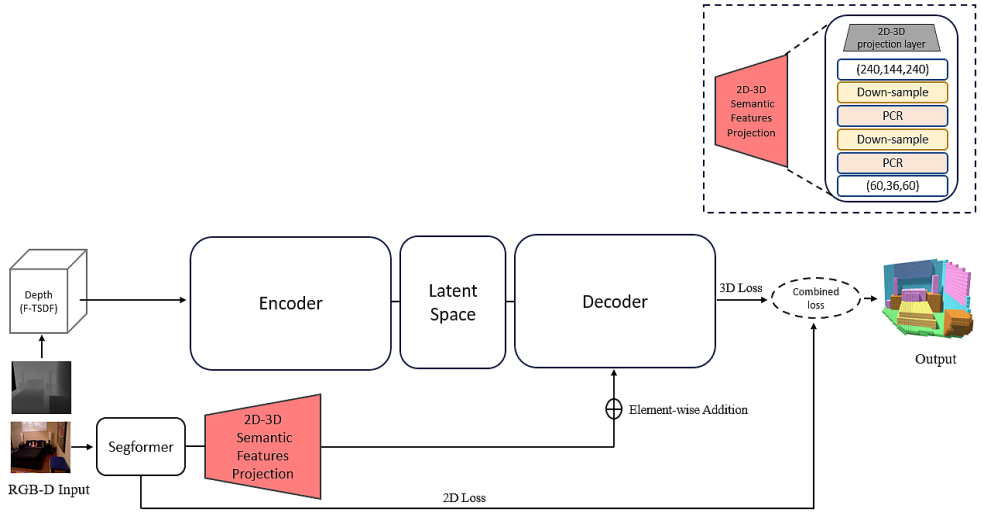

FIGURE 5.4: Late fusion of RGB semantics features in the network.

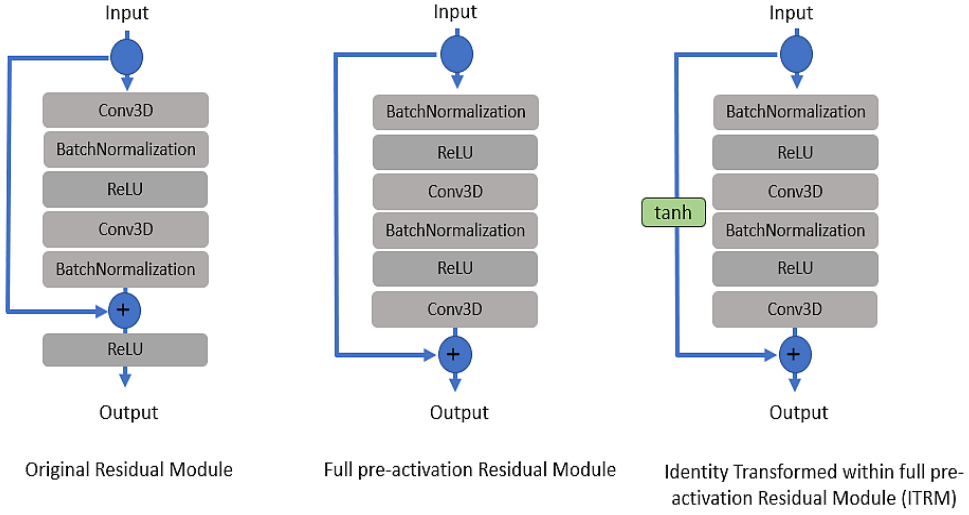

FIGURE 5.5: Different residual block representations. From left to right: (1) Original Residual Module (He et al., 2016b); (2) Full pre-activation Residual Module (He et al., 2016b); (3) Identity Transformed within full pre-activation Residual Module (ITRM), the proposed modification to the full pre-activation residual module.

within a normalized range, aligning with the inherent data range of TSDF or SDF, as demonstrated in (Park et al., 2019; Weder et al., 2020).

In the domain of SSC, the Tanh activation function has been applied to part of identity features, albeit in a different context (Zhang et al., 2019). Our research extends this exploration by investigating additional context for the application of Tanh.

The residual blocks in our model adopt the full pre-activation design outlined in (He et al., 2016b), where batch normalization (BN) and the rectified linear activation function (ReLU) are applied before the convolution layers in a reverse order compared to the standard design, in which BN and ReLU are applied after the convolution layers.

This reversal order facilitates smoother information propagation and performance optimisation. The Equations 5.2 and 5.3:

$$x' = f(BN(x_l)), \tag{5.2}$$

$$x_{l+1} = x_l + F(x', W_l), \tag{5.3}$$

illustrate the relationship between the input and output of the full pre-activated residual block, where the input to the $l - th$ residual block is x_l and the output is x_{l+1}. The function f represents the activation function applied to the normalized input x_l. The residual function $F(x', W_l)$ represents, for example, a series of two convolutional layers, each with a 3×3 filter, applied to x', the pre-activated input in Equation 5.2. The term W_l includes a collection of weights (including biases) associated with the $l - th$ residual block. We modify the full pre-activation residual block design by applying a non-linear transformation with the Tanh function on the identity x_l, as illustrated in Equation 5.4:

$$x_{l+1} = Tanh(x_l) + F(x', W_l). \tag{5.4}$$

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]. Figure 5.5 illustrates the design of the original residual block, the full pre-activation residual block, and the proposed modification within the full pre-activation residual block.

5.2.6 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. The combined loss function is formulated in the following Equation 5.5:

$$L = \lambda L_{SS} + L_{SSC}. \tag{5.5}$$

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} weighted cross entropy loss, defined in Chapter 4, evaluates the model's performance in 3D space, specifically using F-TSDF after integrating projected 2D semantic features in the current context. L_{SSC} combines the benefits of re-sampling and class-sensitive learning

to address the inherent class imbalance in the data. It employs a smoothed weights through an unsupervised clustering algorithm, K-means.

The computation of L_{SSC} loss assesses the discrepancy between the predicted label p and the genuine label y across the voxels of a scene A. For each voxel v within A, the predicted and actual labels for a given voxel v are indicated by p_v and y_v, respectively. Each voxel label is assigned a specific weight w_v using the reweighing method based on K-means clustering. The loss function is defined as follows in Equation 5.6:

$$L_{SSC}(p, y) = -\sum_{v=1}^{A} w_v \,.\, y_v \,.\, log p_v. \tag{5.6}$$

5.3 Implementation Details

The implementation of this work is divided into three main phases: preprocessing, training and validation, and evaluation. The code and processed data can be accessed at:https://github.com/MonaIA1/Repo.

5.3.1 Data Preparation

Similar to the approach described in Chapter 4, we encode all 2D depth maps from the NYUv2 and NYUCAD datasets into 3D space using F-TSDF. The processed data is saved for reuse across multiple designs. We align the 3D scenes layout with Manhattan world assumption, which is related to the direction of gravity. The defined 3D space dimensions are 4.8 meters in width, 2.88 meters in height, and 4.8 meters in depth. With a voxel grid size of 0.02 meters, this configuration results in a volumetric resolution of $240 \times 144 \times 240$ voxels. The TSDF truncation value is set to 0.24 meters, optimizing the balance between detail capture and computational efficiency. Both occluded empty data re-sampling and correspondence between 3D spatial points and 2D RGB pixels using depth maps are established in this stage.

5.3.2 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 MDB-Net. Due to the two types of input representation (the 2D RGB and the 3D geometrical input represented by F-TSDF)we employ different learning rates to achieve effective performance as demonstrated in (Yao and Mihalcea, 2022). Additionally, we adopt different schedulers and optimisers fitted to our network branches contexts. For the 2D input modality (RGB), we employ a pre-trained Segformer model, which is fine-tuned

on the ADE20K dataset at an image resolution of 640×640. The model weights are downloaded from Hugging Face (NVIDIA, 2024). 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 approach suggested by (Wang et al., 2023), we used the AdamW optimizer 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. 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 normalized 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, 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.

5.4 Evaluation

5.4.1 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.

5.4.2 Metrics

We adopt Precision, Recall, and IoU as the evaluation measures for the SSC, following the approach of Song et al. (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's 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., 2019b), DDRNet (Li et al., 2019a), and AICNet (Li et al., 2020a), 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., 2020a), 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.

5.5 Experiments

5.5.1 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.

Fusion Strategies. The model with the proposed combined loss function only is trained using various methods to fuse the 3D projected RGB semantic features as explained in Section 5.2.4. The results, as reflected within average scores presented in Table 5.1, indicate that our model is capable of learning effectively using these different fusion strategies. 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 recognized in some folds when using the early and middle fusion methods, whereas it is consistently recognized 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.

Architecture Components. To confirm the impact of each component within our MDBNet, we modify the DBNet model in the previous Chapter 4 by integrating new components and conduct comprehensive experiments to evaluate their contributions, as detailed in Table 5.2. Initially, we train our model with RGB-D input and apply

TABLE 5.1: Ablation studies using different RGB features fusion methods.

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

TABLE 5.2: 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 combined loss, which includes the re-weighting 3D loss defined in Section 4.2.1, achieving an SSC score of 59.0%.

In the second experiment, we replace the re-weighting loss 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 combined loss in the model's performance.

In the third experiment, we employ our combined loss and enhance the 3D branch of MDBNet by replacing the original residual blocks with the proposed ITRM blocks. This enhancement yields further improvements, achieving an SSC score of 60.1%, a 7.6 pp increase compared to the second experiment's score of 52.5%.

5.5.2 Comparison with State-of-the-Art Methods

Experiments are conducted to evaluate the performance of our proposed approach on scene completion and semantic scene completion tasks, using the NYUv2 and NYUCAD datasets. Quantitative comparisons of our MDBNet results with SOTA approaches are detailed in Tables 5.3 and 5.4. Unlike previous studies, which do not specify the performance uncertainty, we average our scores across three folds to more accurately represent generalization performance. Due to the variations in how researchers select the scene completion area, as discussed in Section 2.3.4, these differences do not necessarily show true performance gaps between SOTA models. Also, (Liu et al., 2024; Li et al., 2020a) highlight the importance of mIoU over IoU. However, for a fair comparison, we focus on semantic scene completion, which relates to the object area and is measured using standardized criteria.

We compare MDBNet with SOTA methods that utilise hybrid architectures, focusing on voxel-based semantic segmentation on the NYUv2 dataset, as shown in Table 5.3.

TABLE 5.3: Results on the NYUv2 dataset include averages and standard deviations for Precision, Recall, IoU, and mIoU metrics. In the input column, 'D' means depth map only. In the method column, '*' represents the view-volume architecture type.

Method	Input	Res.	Scene Completion (SC)			Semantic Scene Completion (SSC)												
			Prec.	Recall	IoU	ceil.	floor	wall	win.	chair	bed	sofa	table	tvs	furn.	objs	mIoU
PALNet (Li et al., 2019b)	D	(240,60)	68.7	85.0	61.3	23.5	92.0	33.0	11.6	20.1	53.9	48.1	16.2	24.2	37.8	14.7	34.1
DBNet (ours in Chapter 4)	D	(240,60)	79.3±0.99	83.3 ±0.84	68.1±0.51	48.9	92.8	49.2	0.0	31.7	61.4	56.1	29.2	0.0	33.9	19.3	38.4±0.17
AMMINet$_{Segformer}$(Wang et al., 2024a)	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
MDBNet (Ours)	RGB-D	(240,60)	80.3±3.7	81.8 ±6.5	67.6±2.1	47.2	92.6	49.9	47.6	46.8	66.2	62.1	37.1	35.7	45.2	36.9	51.6±1.5

TABLE 5.4: Results on the NYUCAD dataset include averages and standard deviations for Precision, Recall, IoU, and mIoU metrics. In the input column, 'D' means depth map only. In the method column, '*' represents the view-volume architecture type.

Method	Input	Res.	Scene Completion (SC)			Semantic Scene Completion (SSC)											
			Prec.	Recall	IoU	ceil.	floor	wall	win.	chair	bed	sofa	table	tvs	furn.	objs	mIoU
PALNet (Li et al., 2019b)	D	(240,60)	87.2	91.7	80.8	54.8	92.8	60.3	15.3	43.1	60.7	59.9	37.6	8.1	48.6	31.7	46.6
DBNet (ours in Chapter 4)	D	(240,60)	86.5±0.91	91.1 ±1.08	79.6±0.13	66.7	93.6	60.7	15.7	51.4	68.9	68.7	45.6	0.0	44.9	29.3	49.6±1.22
AMMNet$_{segformer}$(Wang et al., 2024a)	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
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

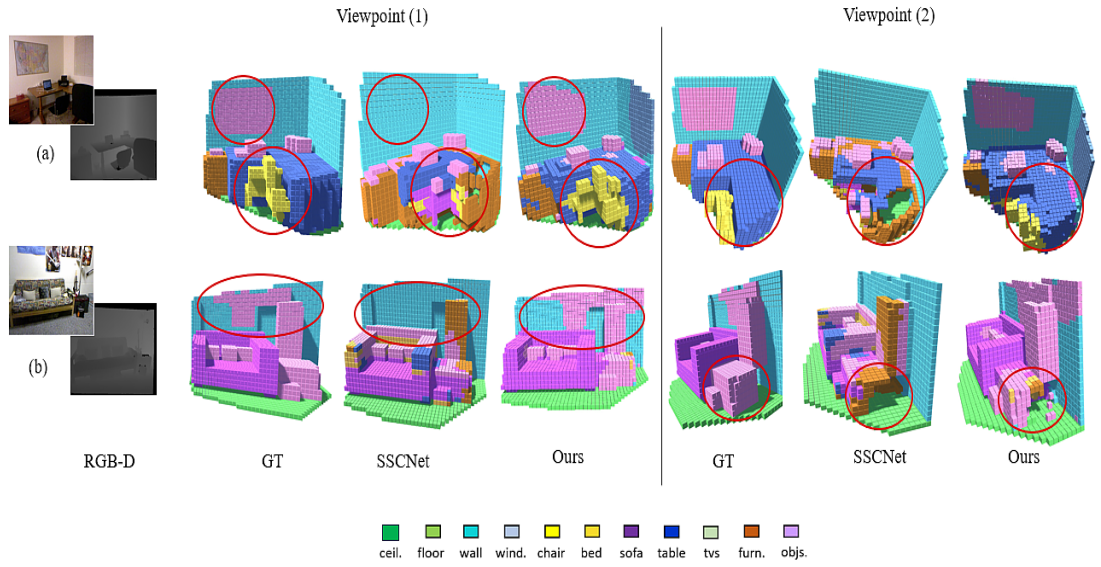

FIGURE 5.6: Comparison of SSC results on the NYUv2 dataset: SSCNet (depth maps) vs. MDBNet (RGB-D). Objects are color-coded, with circles marking key differences between GT and predictions.

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., 2024a) which employs Segformer pre-trained model for 2D RGB features, and PCANet (Li et al., 2023), respectively. This establishes MDBNet as the new benchmark in SOTA performance. The efficacy of MDBNet is further confirmed on the NYUCAD dataset as depicted in Table 5.4. MDBNet shows an increase in the average mIoU scores compared to top previous methods, such as PCANet (Li et al., 2023). Furthermore, although our design surpasses SPAwN on the NYUv2 dataset, it demonstrates performance comparable to the more resource-intensive SPAwN model, which utilises semantics priors calculated using surface normals.

5.5.3 Qualitative Analysis

To highlight the superiority of the MDBNet design and its success in generating more precise predictions, we present a series of visual comparisons using the NYUv2 dataset, as illustrated in Figure 5.6. These comparisons, made between our method and SSCNet (Song et al., 2017), demonstrate the enhanced prediction accuracy offered by our approach. By employing our re-weighting method defined in Section 4.2.1 within the our combined loss and ITRM, we achieve enhanced scene completion, particularly in the occluded parts of the scenes, as demonstrated in (a) and (b) of Figure 5.6. Additionally, by extracting semantic features from the RGB inputs, MDBNet exhibits superior performance, even surpassing the ground truth (GT) 3D volumes in certain regions. For instance, in Figure 5.6 (a), the RGB image shows both object and window existing on the walls. Our model successfully predicts the object and window voxels on

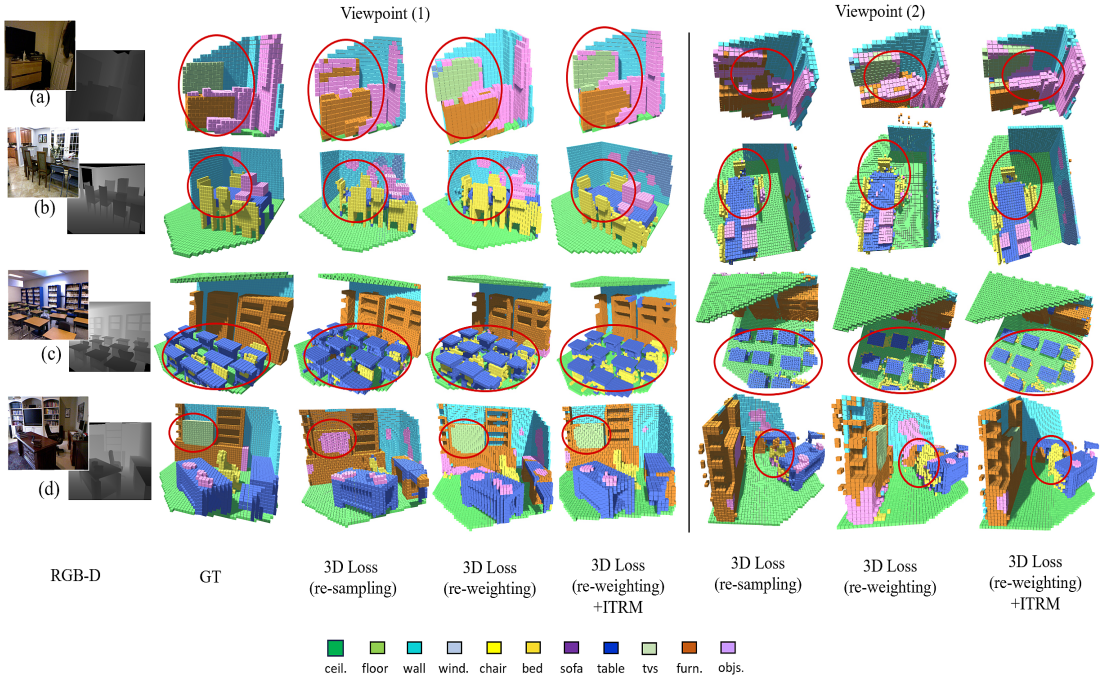

FIGURE 5.7: SSC results with different components on NYUCAD dataset. From left to right: (1) RGB-D input; (2) GT; (3) combined loss with re-sampling; (4) combined loss with re-weighting; (5) combined loss (using re-weighting) with ITRM blocks. Objects are color-coded, with circles highlighting key differences between GT and predictions.

the walls where they are absent in the GT 3D volumes. To illustrate the effectiveness of MDBNet's components, Figure 5.7 showcases various scenarios within the NYU-CAD 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 class re-weighting defined in Section 4.2.1, and when employing our class re-weighting defined in Section 4.2.1 and incorporating ITRM. The incorporation of class re-weighting in our combined loss significantly enhances the model's ability to identify underrepresented classes, such as TVs and chairs, as shown in Figure 5.7 (a), (c), and (d). Additionally, our final design MDBNet offers better recognition of chairs with various shapes in Figure 5.7 (b), (c), and (d), and it ensures enhanced differentiation between tables and chairs, as evident in (b) and (c). MDBNet model effectively recognizes challenging classes like windows and TVs, showcasing its robustness and adaptability.

5.6 Discussion and Summary

In this chapter, we address the 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

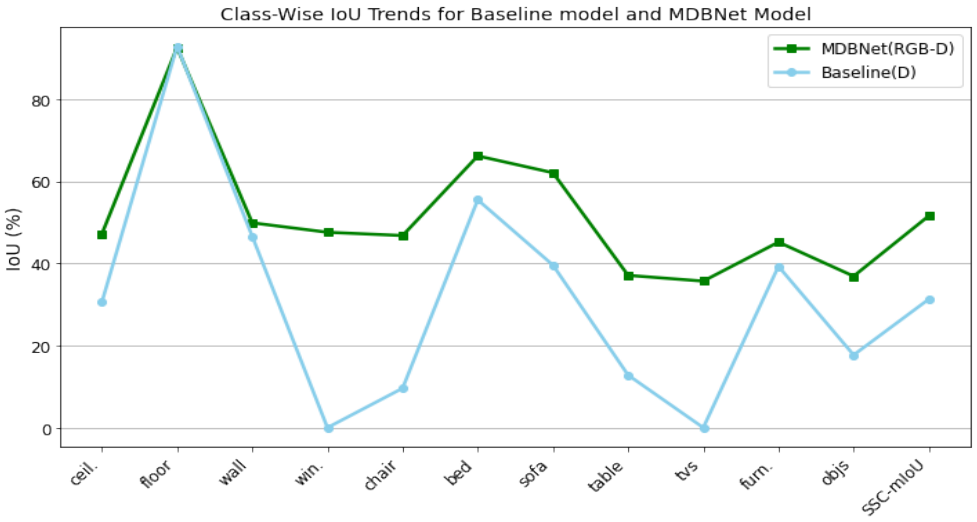

FIGURE 5.8: IoU performance on NYUv2 dataset classes using the baseline model with depth input and MDBNet model with RGB-D input.

training methods such as K-fold cross-validation. We demonstrate improvements in the SSC task on the NYU datasets.

As our research targets 3D spatial modeling suitable for virtual reality (VR) applications, we formulated the following research question:

RQ 2: How can we extend the inference of pre-trained SSC model on perspective images to 360° RGB-D single input?

To address this question, we answer the sub-question related to RQ 2. The first sub-question, RQ 2a, was thoroughly examined in Chapter 4. In this chapter, we focus on addressing the remaining sub-question:

- RQ 2b: What is the impact of learning multiple features from RGB-D input on the performance of the proposed SSC model?

To address RQ 2b, we explore effective strategies for integrating RGB semantic features into the proposed SSC model. Initially, 2D RGB semantic features are extracted using transfer learning, followed by a 2D-to-3D projection approach. Various fusion strategies, including early, middle, and late fusion of RGB semantic features, are systematically examined. 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 5.1, 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. This finding is consistent with prior SSC fusion results reported by (Roldao et al., 2022).

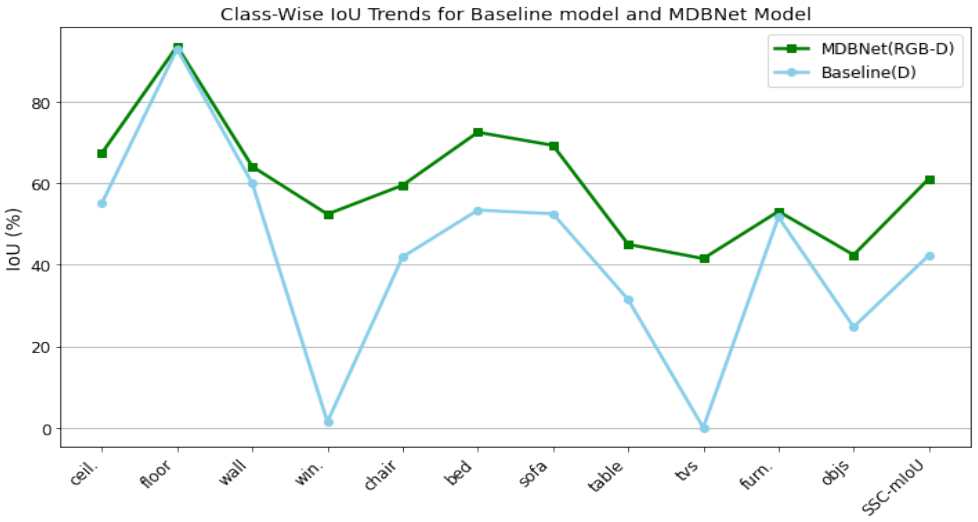

FIGURE 5.9: IoU performance on NYUCAD dataset classes using the baseline model with depth input and MDBNet model with RGB-D input.

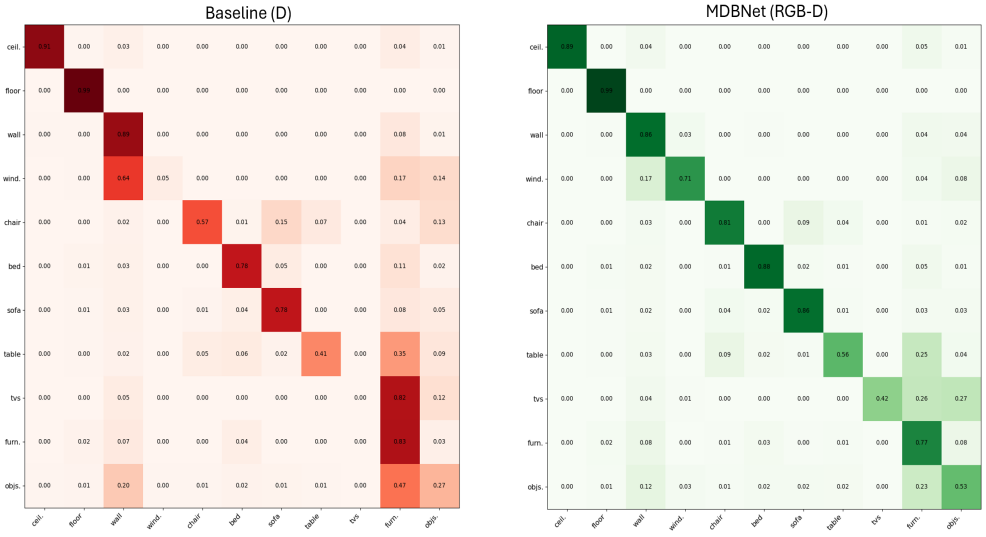

FIGURE 5.10: Confusion matrix for baseline model with depth only input (left) and MDBNet model with RGB-D input (right) on NYUCAD testing set. The diagonal represents correct predictions, while off-diagonal entries indicate misclassifications. The x-axis corresponds to predicted classes, and the y-axis to ground truth classes.

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 5.8 and Figure 5.9 illustrate the baseline, and MDBNet SSC performance over the categories level on NYUv2 and NYUCAD dataset, respectively. The proposed MDBNet model shows a significant improvement in overall mIoU performance compared to the baseline model introduced in Chapter 4. This improvement also highlights the ability of MDBNet to identify challenging object classes, such as TVs and windows, which posed significant difficulties for the previous DBNet design, as discussed in Chapter 4.

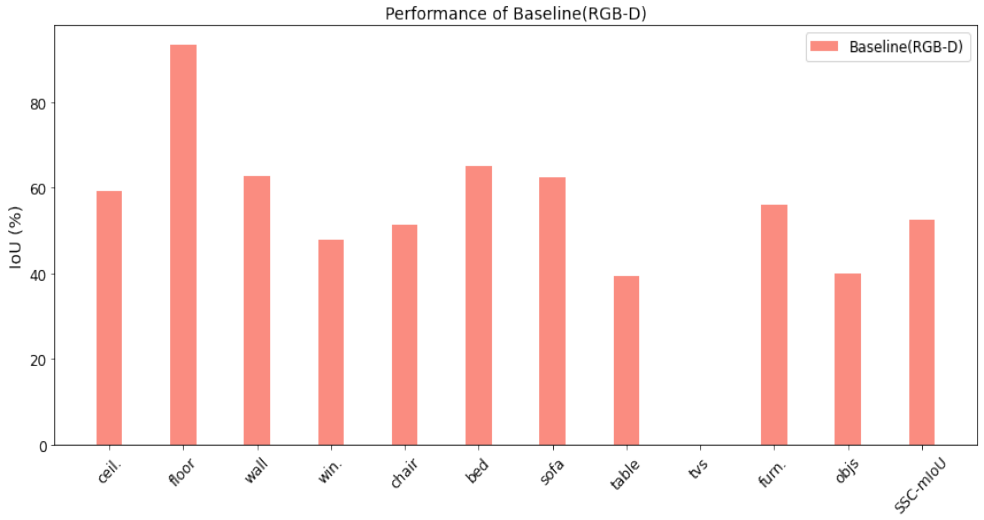

FIGURE 5.11: The baseline architecture performance on semantics level within NYU-CAD dataset.

Additionally, the IoU trends of the MDBNet model compared to the baseline demonstrate enhanced performance within individual object classes, including TVs, beds, windows, furniture, and objects, showcasing its ability to better recognition of single classes. Also, our proposed method shows less ambiguity between classes compared to the baseline. For example, Figure 5.10 demonstrates the confusion matrix of the baseline with depth only input and MDBNet with RGB-D input. It shows that the baseline frequently misclassifies chairs as sofas and confuses windows with walls. These ambiguities are reduced when RGB features are incorporated into the MDBNet model. We observe from the confusion matrix of MDBNet that a general improvement in the diagonal values, which represent true predictions, along with a reduction in off-diagonal values which represent the confusion between classes. The decrease in off-diagonal values indicates a reduction in interclass ambiguity compared to the baseline, which relies solely on depth input.

It is important to highlight that the overall performance gains result from the combined contributions of various components within our design, rather than solely from incorporating RGB semantics. As illustrated in Figure 5.11, adding RGB features without integrating our proposed combined loss function and ITRM blocks 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.

Moreover, MDBNet model is trained on both NYUv2, which includes noisy depth maps from the Kinect sensor, and NYUCAD, which features synthetic and more accurate depth maps aligned with 3D models. We observed that our design reduced the gap

between the SSC scores on NYUv2 and NYUCAD. As shown in Table 5.3, MDBNet achieved an SSC score of 51.6% on NYUv2, compared to 60.1% on NYUCAD in Table 5.4, resulting in a difference of 8.5 pp. This represents a reduction in the performance gap compared to the baseline model with depth-only input, which exhibited an 11 pp gap, with scores of 31.3% on NYUv2 and 42.3% on NYUCAD, as reported in Chapter 4. These results suggest that the proposed MDBNet components not only enhanced overall model performance but also demonstrated increased reliability against real-world noisy depth data.

However, our MDBNet model currently predicts 3D models from perspective camera inputs, which limits its applicability for designing VR spaces with full surroundings. In the next chapter, we extend our MDBNet SSC model to infer 3D models from full panorama RGB-D inputs, enabling the capture of comprehensive surroundings and addressing RQ 2.

Chapter 6

MDBNet 360°: 3D Scene Reconstruction from a Single 360° Image for Virtual Reality with Acoustics

6.1 Motivation and Contributions

Both visual and synchronised spatial audio are essential for creating truly immersive virtual reality (VR) experiences (Stecker et al., 2018; Privitera et al., 2024; Kim et al., 2020b). The integration of both audio and visual aspects enables users to perceive a digital 3D space that closely mimics real-world environments. However, there is a scarcity of studies that integrate audio and visual cues from a single RGB-D 360° input.

While many studies have advanced the visual aspects of virtual reality (VR), particularly in 3D visualisation and human-machine interaction (Rix et al., 2016; Sun and Saenko, 2014; Spiess et al., 2024; Dang et al., 2023; Ciekanowska et al., 2021; Sabir et al., 2024), there are still few studies focusing on the construction of 3D annotated models from 360° inputs. For instance, (Li et al., 2024) employed CNNs for surface reconstruction, but this approach does not extend to generating annotated 3D models with semantic segmentation.

In the 3D semantic scene completion (SSC) literature for indoor scenes, most studies construct 3D models with semantic annotations from perspective views, which suffer from a limited field of view (Song et al., 2017; Li et al., 2019a; Wang et al., 2023, 2024a), where the constructed 3D models do not cover the full surroundings. Therefore, a gap remains in developing pipelines that integrate RGB-D data to generate 3D models with complete semantic annotations from full panorama inputs.

From the audio perspective, studies such as (Chen et al., 2023; Liang et al., 2023; Majumder et al., 2022; Ratnarajah et al., 2024; Singh et al., 2021) leveraged audio-visual inputs to estimate room impulse responses (RIRs), but they neither estimated 3D models nor analysed the relationships between inferred 3D objects and their semantic properties in relation to the estimated RIRs. Consequently, there remains a gap in applying estimated RIRs to predicted 3D meshes for practical use.

Few studies, such as (Kim et al., 2019, 2022), have integrated both audio and visual aspects to create more realistic and immersive VR experiences. The study in (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, the study in (Kim et al., 2022) employed EdgeNet (Dourado et al., 2021), a 3D SSC deep learning model, to infer scene semantics within 3D space. Once the 3D models are built in these studies, sound is rendered within the scenes to contribute to a coherent, immersive experience by integrating both audio and visuals. However, the study in (Kim et al., 2019) simplified scene objects using block representations, while the work in (Kim et al., 2022) demonstrated densely annotated 3D models from depth-only 360° inputs, which suffered from incomplete object reconstructions in the scenes. This motivates us to explore 3D reconstruction using both RGB and depth inputs with 360° FoV while also capturing the RIRs for spatial sound evaluation to achieve a more immersive VR experience.

In this chapter, we investigate the assumption in Chapter 3 that improving semantic scene completion yields more accurate room geometry estimations, ultimately leading to enhanced predictions of acoustic parameters, particularly EDT and RT60. To conduct this investigation, we address the second and third research questions RQ 2 and RQ 3. Specifically, we extend our previously proposed MDBNet model in Chapter 5, to develop a comprehensive 3D framework that integrates 360° RGB-D input. The proposed framework leverages Unity [1], and the Steam Audio Plug-in [2] for advanced 3D sound spatialisation.

Our approach addresses a gap in the existing 3D SSC literature by introducing a unique methodology for processing 360° RGB-D data. 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. 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. The proposed method can be extended to many recent indoor SSC models pre-trained on perspective RGB-D input. Additionally, we analyse the relationship between scene semantics completion and the

[1] https://unity.com/(accessed in 2024)
[2] https://valvesoftware.github.io/steam-audio/(accessed in 2024)

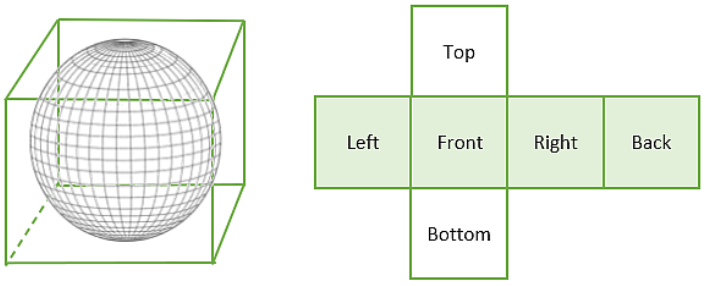

FIGURE 6.1: Spherical to cubic projection

quality of the rendered sound within the full 3D scene by evaluating room impulse response (RIR) acoustic parameters, such as early decay time (EDT) (Barron, 1995) and reverberation time (RT60) (Rungta et al., 2016). We compare our results with state-of-the-art (SOTA) approaches using CVSSP dataset [3].

The summary of our contributions are as follows:

- Propose MDBNet360 a novel method for semantic scene reconstruction and completion leveraging 360° RGB-D input by adapting a pre-trained MDBNet model originally designed for perspective RGB-D data.
- Conduct a comprehensive qualitative assessment to evaluate the reconstructed scenes, and comparing the performance and fidelity with similar SOTA methodological approaches.
- Perform an acoustic analysis of the 3D virtual environments generated by MDBNet360 through the evaluation of RIRs acoustic parameters, specifically examining EDT and RT60, and comparing the results with SOTA methods. The proposed method outperforms SOTA methods showing better 3D scene reconstruction and acoustic parameters for the VR space. The code and rendered sound results are shared via Github account at: `https://github.com/MonaIA1/Repo360`.
- Develop a VR software application in collaboration with graduate students in the School of Electronics and Computer Science (ECS) at University of Southampton to demonstrate the proposed method in a VR setting. The tool can be accessed and downloaded at: `https://chronohaxx.itch.io/avvr`.

[3] `http://3dkim.com/research/VR/index.html`(accessed in 2024)

6.2 3D Reconstruction

6.2.1 SSC with MDBNet.

In the previous Chapter 5, we trained MDBNet on RGB-D perspective camera images from the NYU datasets (Silberman et al., 2012; Firman et al., 2016) to predict scene semantics completion within the 3D space. The model predicts voxel-level semantics for both visible and occluded regions from a single RGB-D partial view. MDBNet's architecture features a dual-head design, enabling simultaneous training with both RGB and depth data encoded using the F-TSDF. This design leverages a combined loss function that integrates 2D semantic loss and 3D loss through a re-weighting scheme specific to SSC tasks. To enhance learning efficiency, the model employs transfer learning to extract 2D semantic features from RGB inputs, projecting these features into 3D space. Depth maps are converted into 3D voxel grids with dimensions of 240 × 144 × 240, where each voxel represents a volumetric unit of two centimeters in size. The RGB features are effectively fused with geometric information derived from depth data encoded as F-TSDF, enabling a more comprehensive scene representation. The model's output has a resolution of 60 × 36 × 60 × 12, where the last dimension corresponds to the 12 semantic class categories present in the scenes. This integration of depth and RGB features enhances the 3D reconstruction process, significantly improving the model's overall performance in understanding and completing scene semantics, outperforming SOTA methods such as (Dourado et al., 2021; Wang et al., 2023; Li et al., 2023).

6.2.2 Extension to MDBNet360.

In this chapter, 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), the spherical RGB image is divided into six perspectives; however, the top and bottom projections, corresponding to the ceiling and floor, are excluded from MDBNet's predictions since they represent known elements that do not require further processing.

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 general principles of spherical-to-Cartesian transformation, as implemented in prior works such as (Kim et al., 2022). For the equirectangular image, the angular size for each pixel (x, y) is defined with unit height equal to

1/image_height and unit width equal to 2/image_width. Using these values, latitude and longitude are computed. The Cartesian coordinates (x, y, z) are then calculated.

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. The transformation of the Cartesian coordinates (x, y, z) for each view is performed using the following rotation matrix:

$$\begin{bmatrix} x' \\ y' \\ z' \end{bmatrix} = \begin{bmatrix} \cos(\theta) & 0 & \sin(\theta) \\ 0 & 1 & 0 \\ -\sin(\theta) & 0 & \cos(\theta) \end{bmatrix} \cdot \begin{bmatrix} x \\ y \\ z \end{bmatrix}. \tag{6.1}$$

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) :

$$F\text{-TSDF} = \text{sign}(TSDF) \cdot (TSDF_{\max} - |TSDF|). \tag{6.2}$$

The sign in Equation 6.2 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 Figure 6.2. 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:

$$O_i = \sum_{j=1}^{m} P_{ij}. \tag{6.3}$$

The class with the highest score is selected to represent the voxel's semantic class within the 3d volume as shown in Equation 6.4:

$$C = \arg\max_i(O_i). \tag{6.4}$$

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

exported to Unity with Steam Audio for object's material assignment and sound rendering.

6.3 RIR Measurement

In this research, we use the Steam Audio plug-in with Unity to render sounds within the 3D volumes exported by MDBNet360 in virtual space. RIR is simulated between a single virtual sound source and a listener and captured by playing ESS audio on the virtual sound source within the 3D scene. To generate the ESS audio, we follow the approach proposed by Farina (Farina, 2007, 2000; Močnik, 2023), utilising Equation 2.1 from Chapter 2:

$$x(t) = \sin\left[\frac{\omega_1 \cdot T}{\ln\left(\frac{\omega_2}{\omega_1}\right)} \cdot \left(e^{\frac{t}{T}\cdot\ln\left(\frac{\omega_2}{\omega_1}\right)} - 1\right)\right]. \tag{6.5}$$

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 frequency ω_2, as depicted in Equations 6.6 and 6.7, respectively. The sweep has a duration of T.

$$\omega_1 = 2 \cdot \pi \cdot f_1/fs \tag{6.6}$$

$$\omega_2 = 2 \cdot \pi \cdot f_2/fs \tag{6.7}$$

We generate an inverse filter and convolve it with the recorded ESS sound, rendered within the 3D scene. The RIR is extracted 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 reverse cumulative trapezoidal integration to assess the decay of the impulse response, followed by 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).

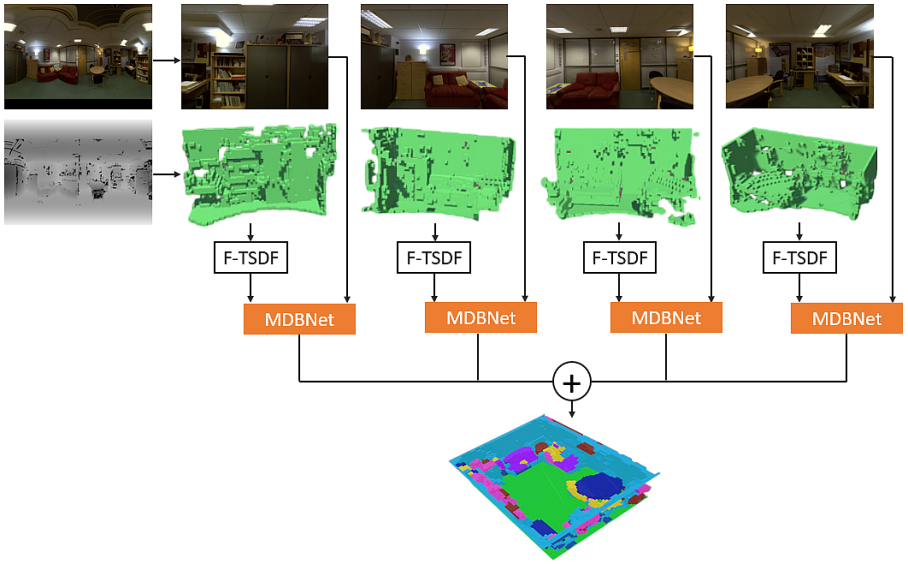

FIGURE 6.2: MDBNet360: RGB-D projection and prediction on full panorama MR scene from CVSSP dataset using MDBNet SSC model.

6.4 Implementation and Experimental Setup

6.4.1 3D Scenes Production.

We test the proposed method using the CVSSP dataset [4]. The CVSSP dataset consists of five indoor scenes with 360° RGB-D and ground-truth acoustic parameter measurements. 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). The SSC model MDBNet utilise a 0.02 meter voxel size within a grid of 240 × 144 × 240 for scene input representation, which is scaled down to 60 × 36 × 60 for the output. For each scene, the camera is simulated to be positioned along the Y-axis and is calibrated to be at scene's center. The 3D predicted volumes are generated by preprocessing RGB spherical images to produce cubic perspective views, combined with F-TSDF 3D data computed using the method described in Section 6.2, with a truncation value set to 0.24 meters.

To infer the 3D volumes, we utilise the saved weights of pre-trained MDBNet model on the NYUCAD dataset. The average inference time to produce a full 3D room is 2.57 minuets on a single Nvidia RTX 8000 GPU. The 3D rooms are exported to Unity

[4] http://3dkim.com/research/VR/index.html(accessed in 2024)

TABLE 6.1: 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

[5] (version: 2022.3.35f1), which is integrated with the Steam Audio plug-in [6] (version: 4.5.3) for immersive sound rendering. Figure 6.2 illustrates the MR scene with 360° RGB-D spherical input, demonstrating scene partitions using our proposed method described in Section 6.2, accompanied by a comprehensive SSC 3D model prediction.

6.4.2 Sound Rendering and RIR Extraction.

In each scene within Unity platform, a virtual sound source and listener are positioned to align with the ground-truth locations. Unified simulation settings are applied across all the scenes. For instance, a corresponding Steam Audio Geometry material is mapped for each object. Table 6.1 lists the objects and their corresponding materials (Kim et al., 2022). Before rendering the sound, the scene must be saved and exported to ensure that all effects, including the geometry materials applied to each component, are correctly integrated. Following the ground truth, where both the sound source and listener are static, we design the simulations using static settings with precomputed, or "baked" effects to reduce CPU usage. An empty game object is added to each scene to assign the Steam Audio Probe Batch, which creates sound probes. These probes serve as points where Steam Audio calculates 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 by using the method described in Section 6.3. The ESS audio generated with a sampling rate of 48,000 Hz and saved at a 16-bit depth. The ESS audio with frequencies ranging from 20 Hz to 20,000 Hz, is rendered with Steam Audio geometry materials within each virtual room. The sound source parameters are illustrated in Figure 6.3.

[5] https://unity.com/(accessed in 2024)
[6] https://valvesoftware.github.io/steam-audio/(accessed in 2024)

To generate spatialize sound, we choose the spatialize option and set the Spatial Blend to the 3D to generate immersive rendered sound as shown in Figure 6.3. 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 involving dynamic sound sources or listeners, which enhances the immersive sound experience. 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 source 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 scene and record the rendered ESS sound. 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 six octave bands as descried in Section 6.3.

6.5 Results Analysis and Comparison with SOTA

6.5.1 3D SSC of 360° Scenes

The original MDBNet demonstrated superior results, significantly outperforming other SSC models, such as EdgeNet (Dourado et al., 2021), FFNet (Wang et al., 2022), and Cleaners (Wang et al., 2023). Due to the lack of ground truth 3D annotated data within CVSSP, we qualitatively assess the 3D voxelized 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. We can clearly observe that MDBNet360 outperforms EdgeNet360 in semantic scene completion across all selected scenes from the CVSSP dataset.

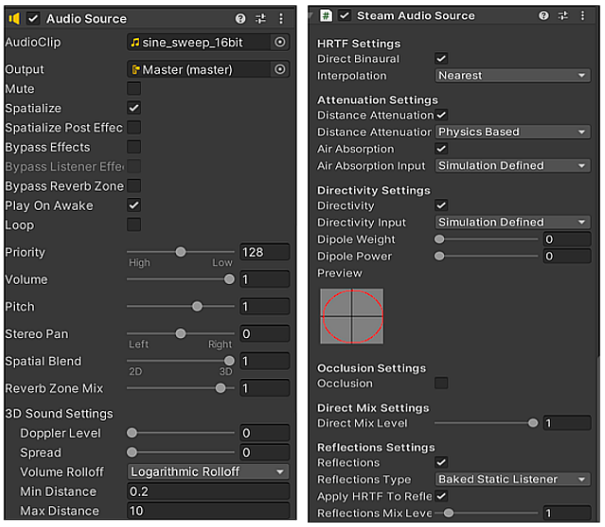

FIGURE 6.3: Sound source settings

Notably, even with the low resolution of depth maps in the CVSSP dataset, where depth values are stored with 8-bit, which leads to a loss of fine object details, MDBNet360 exhibits a clear improvement in predicting and completing key scene components. For evaluation, we focus on objects that play a central role in understanding room structure and functionality, namely sofas, chairs, and tables. These elements were chosen because they are among the most commonly used indoor objects and influence spatial perception. To provide our qualitative comparison, we select a viewpoint that prominently displays these key objects, ensuring a clear visualisation of the model's reconstruction capabilities. As illustrated in Figure 6.4, 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

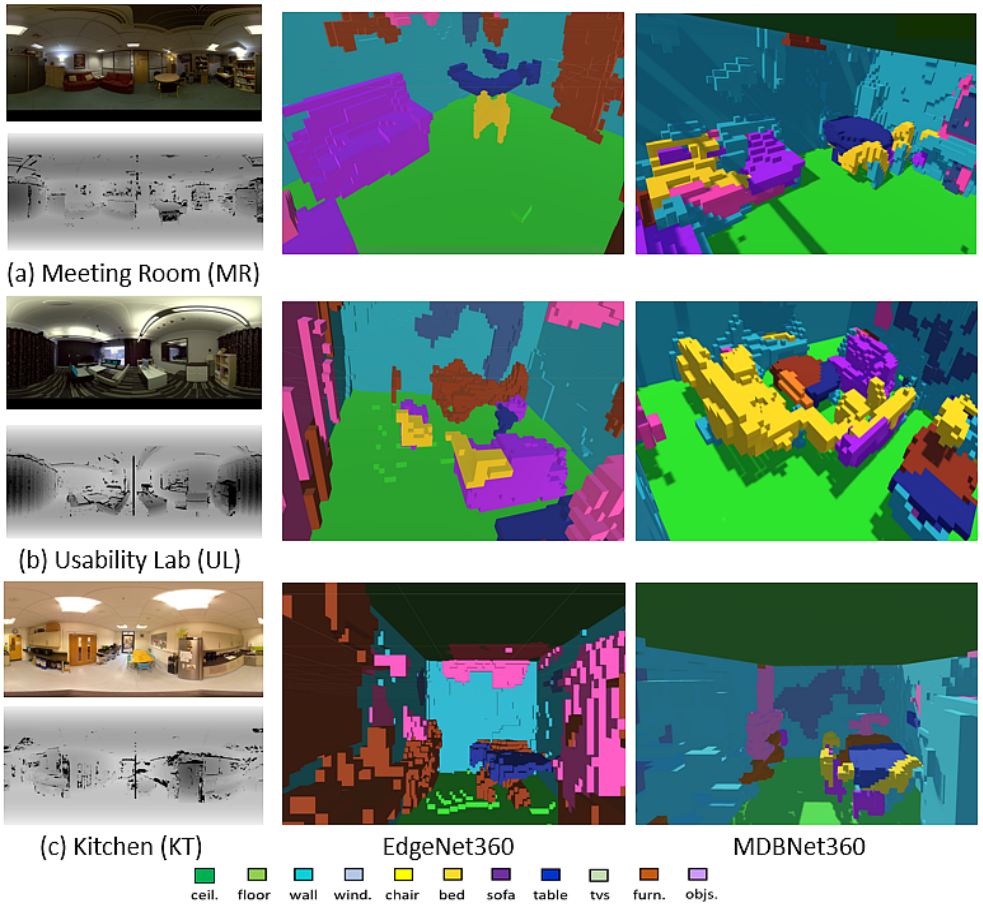

FIGURE 6.4: Qualitative comparison between MDBNet360 and EdgeNet360 on three scenes in CVSSP data. From top to bottom: MR, UL, and KT.

data. Nonetheless, our results indicate that MDBNet360 improves the completeness and fidelity of indoor scene reconstruction, particularly in the representation of essential structural elements which is an aspect crucial for high-quality semantic scene completion.

6.5.2 Spatial Audio within VR Space

To provide a comprehensive evaluation of the VR 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 (for RIR visualisations, refer to Appendix A). Our results are compared with the ground truth measurements obtained from sound modeled in real space, and SOTA models Kim19 (Kim et al., 2019) and Kim20 (Kim et al., 2022). Overall, our approach demonstrates superior performance in both EDT and RT60 compared to Kim19 and Kim20, as shown in Figure 6.5 and Figure 6.6. In Figure 6.5, 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

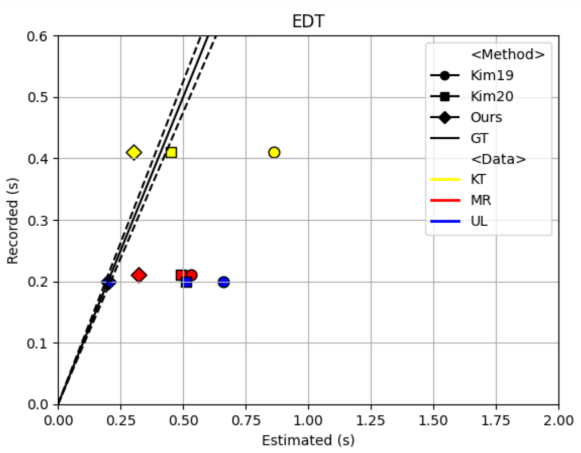

FIGURE 6.5: EDTs for three CVSSP rooms related to the ground-truth (GT).

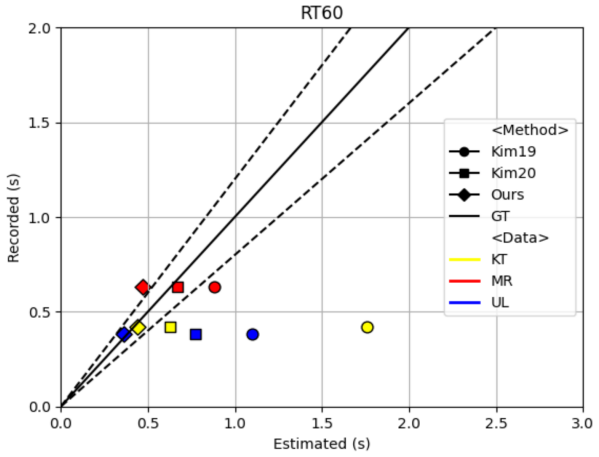

FIGURE 6.6: RT60s for 3 CVSSP rooms related to the ground-truth (GT).

attribute this discrepancy to errors in the 3D semantics, where cabinets are mislabeled as wall voxels. Therefore, plaster materials assigned to cabinets. This mislabeling likely occurred due to inaccuracies in depth perception and the similarity between the cabinet color and the wall color 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 plaster-like material properties.

Additionally, we observe some artifacts that affected the acoustic modeling, 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., 2020b). This can be

FIGURE 6.7: HP Reverb G2 headset with hand controllers connected to the VR application.

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, MDBNet360 demonstrates improved performance in both 3D visual scene prediction and spatial sound rendering compared to existing approaches. The block-based method in (Kim et al., 2019) showed overestimated reverberations due to its simplified, flat surface representations. Similarly, EdgeNet360's reconstructions (Kim et al., 2020b) 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. As discussed in Section 2.5, discontinuities and gaps in reconstructed mesh surfaces negatively impact the sound waves reflections. In contrast, our method preserves scene fidelity through complete object reconstruction as shown across the tested scenes. For example, the geometric precision is demonstrated by more accurate reconstruction of tables and chairs around the tables in both MR and KT scenes. By preserving structural elements, our method achieves more accurate 3D spatial audio. The rendered sound results are shared via Github account at: `https://github.com/MonaIA1/Repo360`.

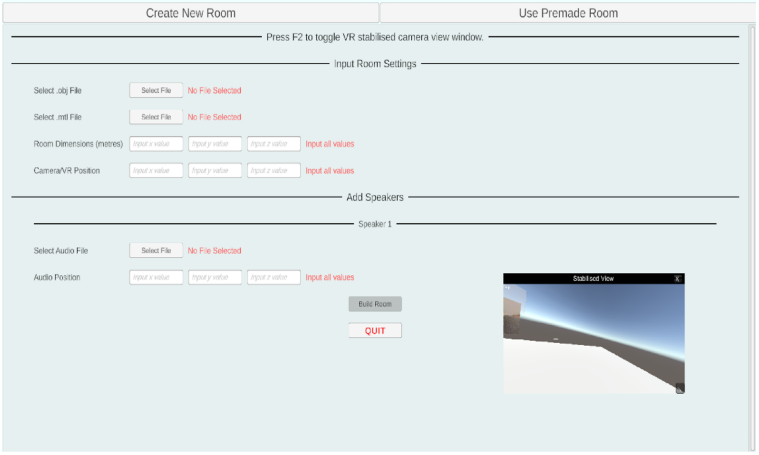

FIGURE 6.8: VR application interface allows users to define and build a custom virtual environment.

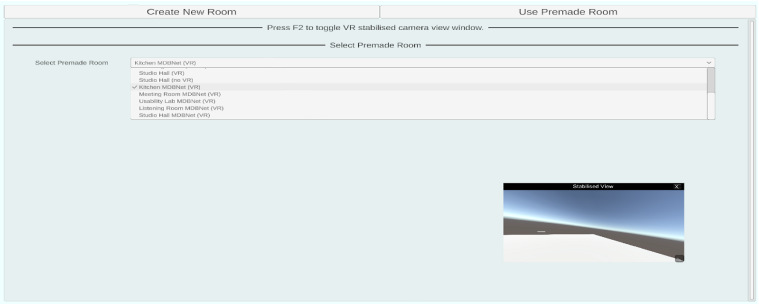

FIGURE 6.9: VR application interface allows users to select premade rooms.

6.6 Real-Time VR Application

In this section we demonstrate the design of a real-time VR application using our proposed method and CVSSP data. This work is done in collaboration with graduate students in ECS. The application designed based on the pipeline represented in Chapter 3, with a modification of the pipeline components using the proposed MDBNet 360° in this chapter with stereo depth maps. Additionally, the pipeline updated with DBAT material recognition model(Heng et al., 2023). The sounds rendered in real-time to provide immersive experience to the users and changes based on users movements around the sound source in the scene. For the VR demonstration, we use the HP Reverb G2 headset with controllers to manage movement and user options in the VR menu, as illustrated in Figure 6.7.

6.6.1 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 organ-

ised into two primary tabs: Create New Room and Use Premade Room, as illustrated in Figure 6.8. These tabs cater to different user needs, allowing users to either construct customized room environments or select and utilise predefined ones. Premade rooms demonstrate the reconstructed 3D models with 360° surroundings.

Create New Room Tab. This tab allows users to construct a custom virtual room environment. The interface is organised into distinct sections that systematically guide users through the room setup process:

- Input Room Settings: Users can upload the required 3D models and the corresponding materials represented by `.obj` and `.mtl` files to specify the geometry and material properties of the room. Additionally, fields are provided for defining room dimensions (in meters) and setting the camera or VR position. These parameters ensure precise spatial configuration for rendering and simulation purposes.
- Add Speakers: The GUI enables the integration of sound source or speaker within the environment. Users can add or remove speakers, with each requiring an audio file upload and the specification of positional coordinates. This modular approach allowing flexible placement and configuration of sound sources.
- Build Room: This to finalizes the configuration and initiates the room's construction.

Use Premade Room Tab. The tab provides a streamlined workflow for users working with predefined 3D rooms. Users can select specific scene, such as a kitchen, as shown in Figure 6.9. The selected scene is loaded by clicking the Load Scene button, minimizing setup time for recurring tasks and common use cases.

6.6.2 XR Interaction Toolkit

The VR application is built using XR Interaction Toolkit 3.0.3 [7] [8]. 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 physical contact and ray-based (distance-based) interactions.

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

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

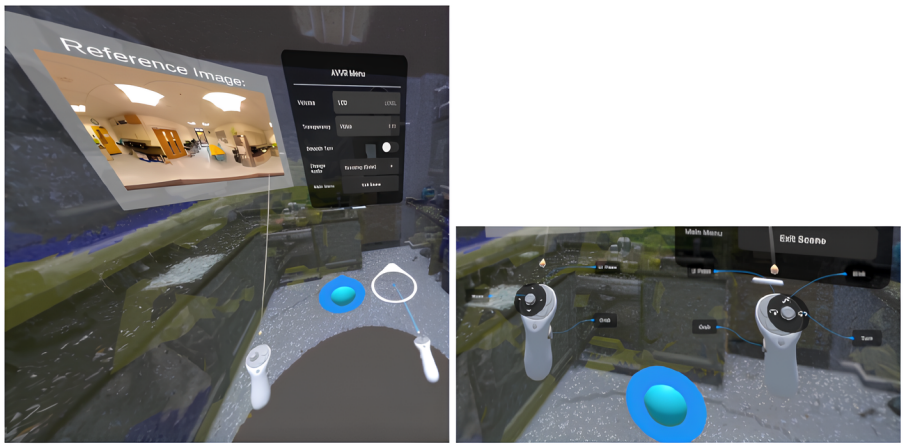

FIGURE 6.10: VR locomotion system showing smooth movement option on the Features menu and the controllers with teleportation.

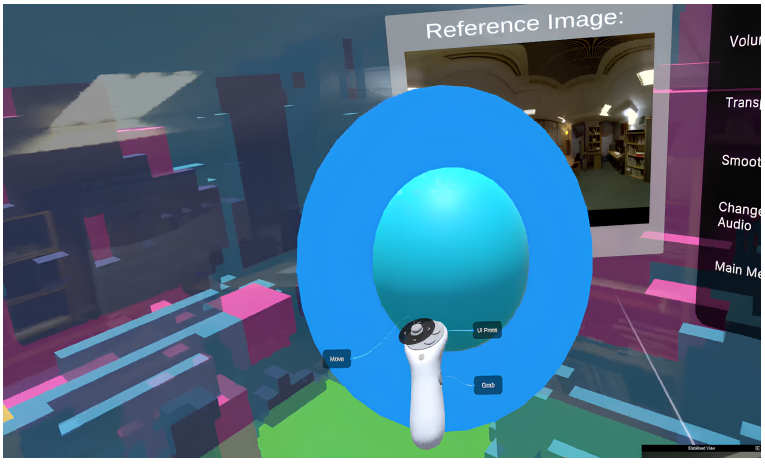

FIGURE 6.11: Illustration of grabbing sound source sphere object (blue) within MR scene.

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 field-of-view (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 6.10 illustrates the controllers and teleportation in VR.

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 also show that the user can interact with the audio source sphere in the

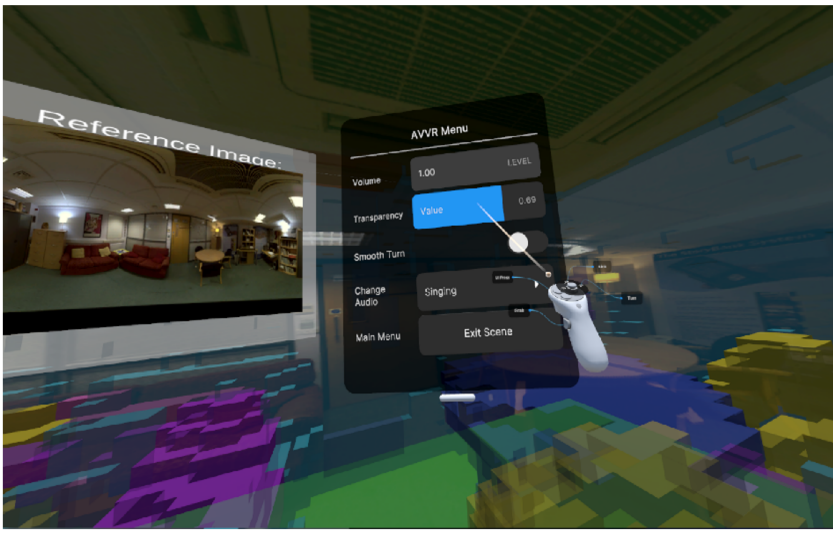

FIGURE 6.12: VR menu showing volume and objects transparency sliders in MR scene.

VR space as a key example of a grabbable object implementation. Figure 6.11 shows the controller holding the audio source.

6.6.3 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.

Audio Volume and Mesh Transparency Controls. Real-time adjustments of both audio levels and mesh visibility are managed through intuitive slider controls as illustrated in Figure 6.12. 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.

Movement and Audio Options. Users can toggle between snap turning and smooth turning based on their comfort preferences, with snap turning providing fixed-angle rotations to reduce motion sickness. The audio system supports multiple options, including music and speech samples at varying volumes as illustrated in Figure 6.13. The audio source maintains proper spatial audio properties with position-based attenuation, enhancing the overall immersion of the experience.

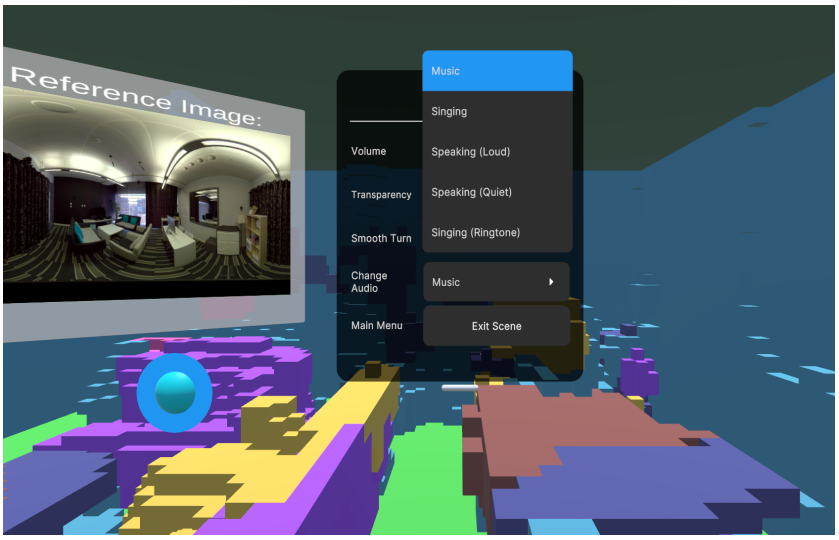

FIGURE 6.13: VR menu showing the audio options in UL scene.

6.6.4 LiDAR Scan Integration

A high-precision LIDAR scan is integrated into the application to enhance the exeriance with the objects texture and presence within the VR space. The LiDAR implementation incorporates appropriate material assignments for acoustic properties to maintain consistent audio behavior. Figure 6.14 illustrate the LiDAR integration to KT scene while keeping the reconstructed 3D model transparent.

6.6.5 VR Application Evaluation and Observations

The proposed VR application is designed to facilitate a high level of interaction for users who are not familiar with the Unity platform. The core functionality of this application is to provide the experience of being inside reconstructed virtual rooms integrated with spatial sounds. An extensive user study was not conducted to evaluate MDBNet360 with real-time rendered sound, as this is outside the scope of the current research; the VR application is presented solely for demonstration purposes. However, we present our observations based on real-time testing of the VR application with the reconstructed SSC rooms in this chapter.

The evaluation focuses on two main aspects: (1) the functionality of the VR menu and available interaction options, and (2) user interaction with the VR environment, including sound sources and spatial audio rendering.

- **VR Menu and Interaction Options:** All implemented menu options function as expected. Users can adjust audio volume through VR controllers and modify the transparency of reconstructed meshes in real-time. For MR and UL scenes, RGB

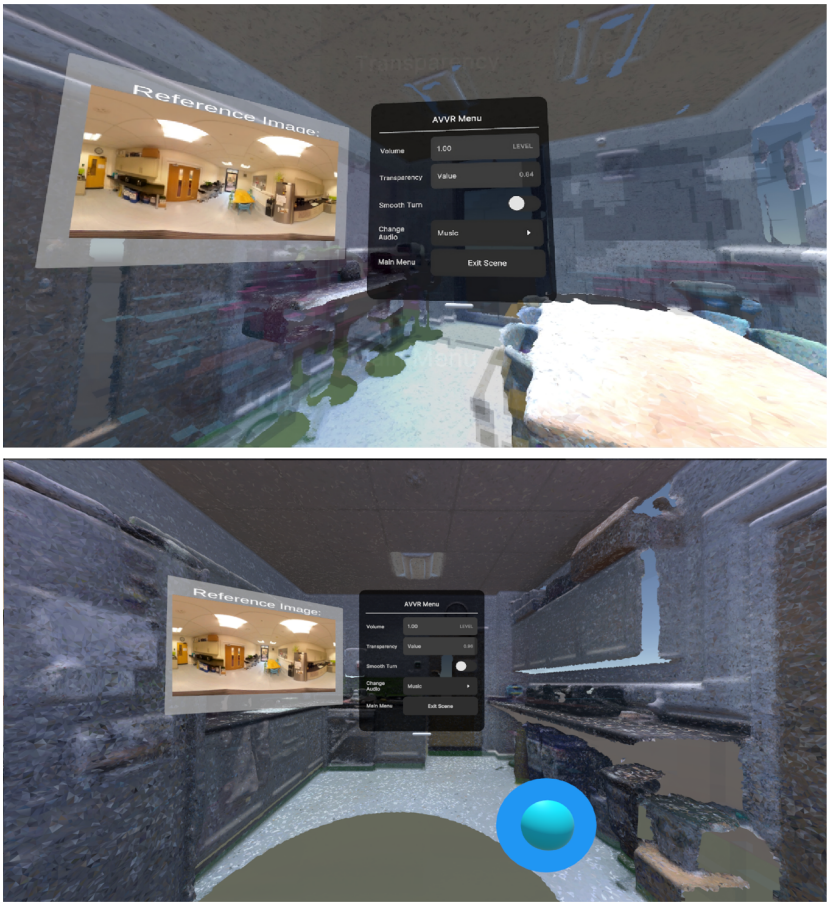

FIGURE 6.14: Illustration of the KT scene with an overlaid LiDAR scan with two different view points.

textures are used as visual references, whereas LiDAR scans are used for the KT scene. These references help users better understand the scene components and enhance the perceived realism of the reconstructed environments. We observe that the LiDAR scan provides detailed textures of the scene, which increases immersion and serves as a spatial reference for the scene components. Furthermore, users can personalise their movement preferences by enabling or disabling the smooth-turn option, which helps mitigate motion sickness. The application also allows switching of audio sources via a dropdown menu. Moreover, users can easily exit a scene and return to the main menu to load another room if desired.

- **User Interaction and Spatial Audio:** The VR 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. 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. 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.

At this stage, 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 showcasing the application's functionality is available at: `https://github.com/MonaIA1/Repo360`.

6.7 Discussion and Summary

In this chapter, we present a method for generating a 3D space for VR, integrating both visual and acoustic cues using a single 360° RGB-D input. To achieve this, we develop the MDBNet360 model, that designed to produce a comprehensive 3D voxelized representation of scenes in virtual space. Our approach is built on the pre-trained SSC MDBNet model introduced in Chapter 5.

We formulate the second research question, RQ2, as follows:

- RQ 2: How can the inference capabilities of pre-trained SSC model on perspective images be extended to a single 360° RGB-D input?

To address this research question, we propose a method that leverages both RGB and depth data through a series of processing steps as described in Section 6.2.2. First, 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. Next, we perform a 3D rotation on point clouds generated from the spherical depth information to ensure proper alignment with the cubic RGB views. To capture the geometric structure, we calculate the F-TSDF for each cubic face. The processed views are then fed into the MDBNet360 model, an extension of the MDBNet architecture design specifically to handle perspective RGB-D inputs, as illustrated in Figure 6.2. The outputs from all 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.

To evaluate the effectiveness of this approach, we conduct a qualitative comparison of the generated models with those produced by EdgeNet360 using the CVSSP dataset. As detailed in Section 6.5.1, our findings demonstrate that MDBNet360 achieved better results, producing more realistic scene reconstructions and improved semantic completion, ultimately enhancing the understanding of the room's spatial structure and functional layout.

Similarly, we answer the third research question, RQ3, which is defined as follows:

- RQ3: What is the impact of the 3D scene generated from a 360° RGB-D input on acoustic parameters, including early reflections and late reverberations?

We evaluate the acoustic quality of the rendered sound within the reconstructed 3D virtual rooms generated by the proposed MDBNet360 model on CVSSP dataset. Specifically, we measure the EDT and RT60 acoustic parameters, which are commonly used to characterize early reflections and late reverberations, respectively.

The evaluation process began by capturing the RIR for each reconstructed room using the ESS method. This method varies frequencies exponentially over time, enabling the effective capture of sound across different frequency bands. To ensure consistency with ground truth data and state-of-the-art models, including those presented in (Kim et al., 2019, 2022), we compute average scores for EDT and RT60 across six octave bands ranging from 250 Hz to 8000 Hz.

Our results demonstrates that the 3D scenes generated by the proposed MDBNet360 produce better EDT and RT60 values compared to state-of-the-art models (Kim et al., 2019, 2022), as discussed in Section 6.5.2 and illustrated in Figures 6.5 and 6.6. Specifically, our model's improved scene semantic completion contributes to closer alignment with ground truth acoustic parameters.

Overall, the findings confirm that the proposed SSC model not only improves the visual semantics of 3D scenes, but also enhances realism of acoustic modeling, 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 more realistic and interactive VR environment by single 360° RGB-D input.

Chapter 7

Conclusions and Future Work

In this thesis, we investigated a challenging task: the generation of a 3D digital space for virtual reality (VR), integrating both visually accurate and acoustically plausible audio from a single 360° input for an indoor scene. Throughout the work, we progressed through various stages, employing different methods and tools to construct and assess both the audio and visual components. In each chapter, we highlighted our contributions and advancements, addressing key challenges and proposing novel approaches for creating immersive, multi-sensory virtual environments.

7.1 Conclusions

In Chapter 3, we demonstrated the generation of a VR space with sound rendering from a single RGB 360° input. Our proposed pipeline integrates estimated mono depth, 3D scene semantics, and material properties, enabling sound rendering within the Unity VR gaming engine using the Steam Audio spatial sound plug-in. Preliminary results have set a roadmap for the research areas in this PhD project by identifying the challenges and potential improvements. A primary focus for enhancement was on the SSC component, which is fundamental to generating a comprehensive 3D model encompassing scene semantics. We proposed to replace the semantic scene completion (SSC) part within the proposed pipeline with a more advanced SSC approach, leveraging RGB features in conjunction with depth data to improve 3D reconstruction accuracy.

In Chapter 4, we addressed the problem of SSC, which involves inferring volumetric occupancy and object categories from a partial depth input. Our SSC model, using a single depth input encoded with F-TSDF for scene representation, predicts complete 3D scenes. We tackled a key challenge in the domain, such as imbalances in 3D data distribution, particularly in voxelized indoor scenes. To address these, we proposed a novel

re-weighting method integrated into our loss function. This method uniquely combines two class re-balancing approaches (re-sampling and class-sensitive learning) and smooths the weights using an unsupervised clustering algorithm, prioritising voxel weights and enhancing the learning process. We employed benchmark training methods such as K-fold cross-validation. Our results outperformed baseline models, particularly in completing scene semantics, on benchmark datasets such as NYU datasets. However, upon deeper analysis of our results at the class category level, challenges remain in recognizing certain underrepresented objects, such as TVs and windows. Identifying these objects from depth features alone is challenging due to the reflective or transparent surfaces of windows and the misclassification of TVs with other object classes sharing similar visual features. This limitation highlights the challenges of using depth information alone in our results.

In Chapter 5, we extended the work in Chapter 4 by proposing the MDBNet model with RGB-D inputs. Our method improved SSC learning by incorporating key components such as a combined loss function, ITRM blocks, and various RGB feature fusion methods, evaluated through K-fold cross-validation. We demonstrated improvements in SSC performance on the NYU datasets. Our analysis of SSC scores revealed that prioritising voxel weights with our re-weighting method, integrated into the combined loss function, significantly enhances SSC learning, particularly for RGB-D inputs. We also found that the late fusion strategy was the most effective for integrating RGB features into the network. Moreover, we showed that incorporating RGB features improves class identification, although challenging scenarios (such as underrepresented classes like TVs) still posed difficulties if not combined with a solid methodology. The power of our SSC predictions stems from the combination of multiple feature inputs (RGB-D) and the proposed components within the architecture.

In Chapter 6, we introduced MDBNet360, a new methodology for generating comprehensive 3D voxelized representations including scene semantics of 360° RGB-D scenes. Building upon the semantic scene completion MDBNet model in Chapter 5, our approach applied a spherical-to-cubic projection to the RGB data, along with a 3D rotation on point clouds from depth information, facilitating the construction of a detailed 3D model that captures the full 360° spatial context. To validate the acoustic quality of our generated scenes, we conducted a comprehensive evaluation by rendering ESS mono sound and analysing room impulse responses (RIRs) parameters. Specifically, we measured early decay time (EDT) and reverberation time (RT60), to assess the acoustic realism and spatial quality of the synthesised virtual environments. The empirical results demonstrate that our proposed methodology outperforms existing state-of-the-art approaches by effectively integrating visual and acoustic cues through advanced techniques.

7.2 Limitations and Future work

Creating a 3D representation of a real-world space that integrates visually accurate details with acoustically plausible audio from a single RGB-D 360° indoor scene is a complex and largely underexplored challenge. While this research contributes to understanding both perspective and 360° camera inputs, several critical open questions and promising avenues for future work remain.

7.2.1 Investigate Advanced Projection of 360° Inputs

In this research, spherical-to-cubic projection was applied to the RGB inputs, along with a 3D rotation on point clouds from depth information, facilitating the construction of a detailed 3D model that captures the full 360° spatial context. It contributes to the SSC domain by introducing a foundational method for aligning 360° RGB and depth data. However, the cubic projection introduces some distortions near the edges of cube faces, and may not generalised for different room shapes. Future work could explore more advanced and perceptually consistent projection methods to address these limitations.

One promising direction is the segmentation of the spherical domain using regular polyhedra, such as the icosahedron or dodecahedron, which provide more faces spread across the sphere and preserve perspective projection (Lee et al., 2019). These structures may allow for more uniform sampling of the scene, reducing edge distortion while improving geometric continuity. Furthermore, improving the alignment and consistency of projected 3D point clouds with these projection methods should be systematically investigated. Advancing projection methods in this direction may enhance the spatial fidelity of 360° data representations and offer better generalization to different room configurations.

7.2.2 Multi-Scale Fusion Architecture

In our SSC MDBNet model, we systematically explored different feature integration strategies using element-wise addition. We found that the late fusion approach is the most effective method for integrating RGB features with 3D geometric data. Our comprehensive investigation considered three distinct fusion methodologies such as early, middle, and late fusion, and demonstrated that late fusion yields the highest SSC performance scores. While the fusion methods proposed in this research proved to be effective, future work could explore more advanced strategies, particularly in multi-scale fusion of F-TSDF data. Inspired by recent research (Wang et al., 2024a), incorporating multi-scale fusion could enable more refined geometric representations and potentially improve SSC performance.

Furthermore, we showed a context of employing Tanh activation function for identity features within the proposed ITRM block. However, this approach merits further exploration. Future research could examine the application of the Tanh activation within cross-modal architectures and TSDF inputs. Such investigations could provide insights into optimizing the use of Tanh for inputs like TSDF, potentially enhancing model performance.

7.2.3 Uncertainty Quantification

Although our SSC MDBNet model demonstrates competitive performance compared to SOTA methods, it faces challenges related to evaluation consistency. The selection of a single evaluation run often introduces bias, as it tends to favor the best-performing instance. To improve performance consistency and address variability, we introduced performance uncertainty quantification through K-fold cross-validation on SSC IoU and mIoU overall scores. However, the lack of established best practices for training and validation in this field remains a challenge. This can be attributed to data nature with high dimensionality used in this task.

Future research could focus on exploring model confidence and calibration techniques suitable for this ill-posed problem, which have the potential to further enhance model performance and reliability. A fundamental challenge in SSC arises from the partial-view nature of its inputs, which inherently results in the loss of 3D information in occluded regions. This missing data, attributed to aleatoric uncertainty (Kendall and Gal, 2017; Kendall et al., 2018), complicates the prediction of volumetric occupancy in these occluded areas. Addressing this uncertainty can improve the SSC performance accuracy. While our research primarily focuses on VR space design including both audio and visual data, time constraints prevented us from exploring this research direction. Future studies might consider exploring uncertainty in SSC indoor scenes as one possible area for refining performance and expanding the understanding of SSC problem.

7.2.4 Generalisation with Audio-Visual 360° RGB-D Datasets

In this study, we encountered a challenge with the limited availability of datasets that provide both RIRs and 360° RGB-D scenes. While datasets such as Matterport (Chang et al., 2017) and 2D-3D Stanford (Armeni et al., 2017) offer high-quality 3D reconstructions, they do not include corresponding RIRs and often require additional post-processing to address gaps or discontinuities in the reconstructed surfaces. For our audio-visual evaluation, we employed the CVSSP dataset owing to its unique integration of RIRs measured within the real space with 360° RGB-D scenes. Our method yielded promising results as shown in Chapter 6; however, we recognise that the CVSSP dataset is limited in terms of generalisation. For example, we observed sound wave

artifacts in one scene, which may stem from inherent limitations of the spatial audio modelling provided by the Steam Audio plug-in. This observation does not detract from the overall performance of our proposed SSC method but rather highlights an opportunity for further research.

Future work could involve expanding the dataset by capturing additional indoor environments with corresponding RIRs and providing 3D ground-truth annotations. This is a nontrivial task 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. For 3D mesh ground truth generation, similar to the NYU dataset (Guo et al., 2015), it would require one to start from pixel semantics to find the objects' regions within the scene. Then, generate 3D CAD models for all objects in each scene with ceiling, walls, and floor layouts, and align them with the corresponding depth maps. This process would require scaling, rotation, and translation to ensure geometric consistency between CAD models and depth data. Capturing high-quality RIRs would also require multiple recordings per scene, testing various microphone placements to ensure robustness in accordance with acoustic measurement guidelines such as ISO 3382-1 (International Organization for Standardization, 2009). This would enhance the generalisability of our method and support the development of more accurate spatial audio plug-ins.

7.2.5 Applying Knowledge Distillation in SSC for VR Space Using Monocular 360° RGB

One important direction in SSC research is predicting 3D volumes from monocular RGB images, as explored in (Cao and de Charette, 2022; Yao et al., 2023). However, this task is challenging due to the lack of geometric depth information in monocular RGB inputs. A promising solution is the use of knowledge distillation. We propose extending our MDBNet method to predict 3D volumes from monocular RGB by leveraging the knowledge distillation framework introduced in (Hinton, 2015). Knowledge distillation transfers knowledge from a more powerful teacher model to a simpler student model.

For instance, (Wang et al., 2023) used knowledge distillation in SSC by training a teacher network on noise-free depth maps (NYUCAD dataset) and distilling its knowledge into a student network trained on noisy depth maps (NYUv2 dataset), helping the student network correct prediction errors. In contrast, a future work of this research involves training a teacher network with RGB-D inputs to leverage geometric information and distilling its knowledge into a student network that uses only RGB inputs. Both features or logits can be distilled from the teacher network (Wang et al., 2023; Ji et al., 2021; Zhao et al., 2022), enabling the student network to generalise effectively for predicting 3D volumes from monocular RGB. This makes the method suitable for real-world applications where depth data is unavailable.

By incorporating geometric insights from the teacher network, we aim to bridge the gap between RGB and RGB-D-based SSC models, addressing the core challenges of monocular RGB-based 3D scene completion. Furthermore, as an exciting extension, we propose adapting and simplify the inference pipeline in our designed VR space to rely solely on 360° RGB input. This research direction holds promise for enhancing VR applications, enabling immersive and efficient 3D scene reconstruction from readily available monocular RGB data.

7.2.6 Optimise the 3D SSC Learning Using Multi-modal Inputs and Other Emerging Deep Learning Trends

In this work, we employed multiple data representations to train the proposed 3D SSC model, using both RGB and depth information for each scene. However, future work could investigate the integration of additional modalities to further enhance 3D semantic understanding. One promising direction is the incorporation of text semantic guidance through vision-language models. Inspired by multi-modal learning in (Rao et al., 2022), contextual cues can be extracted from the input image to generate natural language prompts. These prompts can be processed by a pretrained language model to produce text embeddings, which then serve as semantic priors that support RGB and depth data. Integrating these priors into the 3D SSC architecture could improve the model's ability to predict more accurate 3D scene semantics.

In addition, replacing the UNet backbone used in this research with a diffusion model architecture, as proposed by (Wang et al., 2024a), could further improve the network's ability to recover both semantic and spatial structures. Further enhancement could be achieved by employing transformer-based encoders such as ViTAE (Xu et al., 2021) to process depth inputs as in (Liu et al., 2024), to improve the modeling of both global and local contextual information. Integrating these emerging architectures could significantly boost the effectiveness and generalization of future 3D semantic scene completion models.

7.2.7 Investigating the Generation of VR Space for Outdoor Scenes using Monocular 360° Inputs

The generation of VR spaces for outdoor scenes presents unique characteristics compared to indoor environments, particularly in terms of data representation and acoustic rendering. For example, outdoor datasets such as the KITTI dataset (Liao et al., 2022) differ from indoor datasets like NYU, commonly used in this research. The KITTI dataset employs point cloud representations focused solely on surface geometry, which requires careful consideration during the design and evaluation of the SSC model.

Additionally, sound evaluation metrics diverge between indoor and outdoor environments. In indoor settings, the reverberation time such as RT60 depends heavily on room geometry and surface reflections. Conversely, outdoor acoustic assessments emphasize sound propagation through the air, where decay times are generally shorter and more variable due to the absence of enclosing structures (Mascia et al., 2015). These distinctions underscore the importance of tailoring both visual and acoustic modeling approaches when generating VR spaces for outdoor scenes.

Appendix A

Room Impulse Response Visualisation

The EDT is calculated from the energy decay curves (EDCs) using the range from 0 to -10 dB to estimate the 60 dB decay. In contrast, the RT60 is calculated over the full 60 dB decay range. Both EDT and RT60 are computed in octave subbands from 250 Hz to 8000 Hz. In our experiments, we averaged the scores across these six octave bands for the MR, UL, and KT scenes. Figure A.1, Figure A.2, and Figure A.3 visualise the RIRs and the acoustic parameters EDT and RT60 fitted to the decay curves for the MR, UL, and KT scenes, respectively.

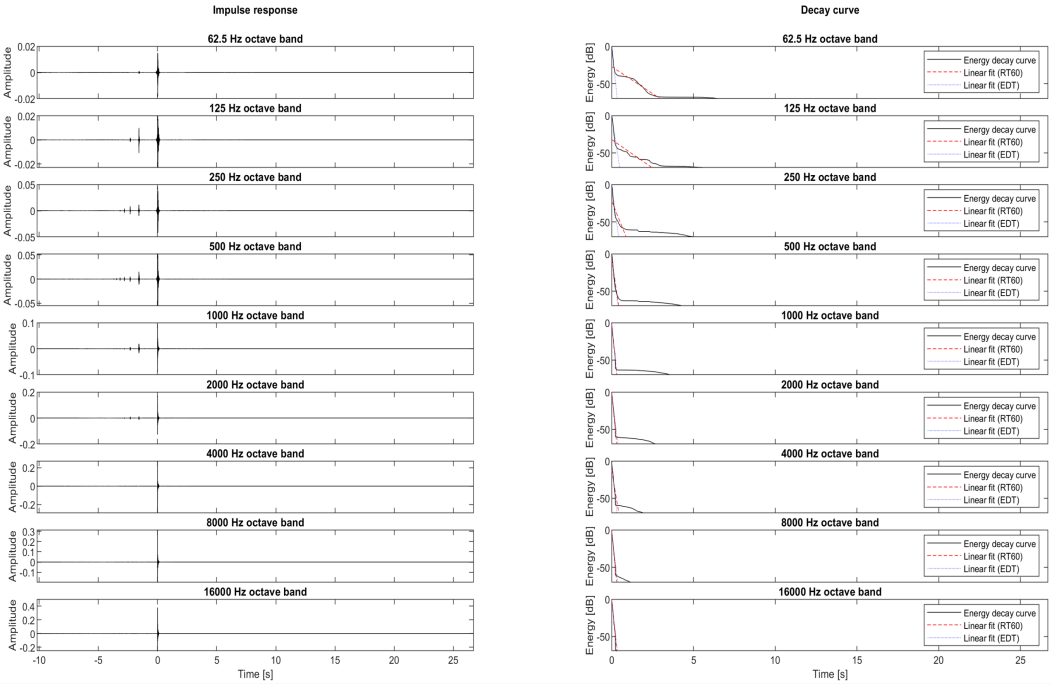

FIGURE A.1: Meeting Room (MR) RIR visualisation and energy decay curves over different octave bands, showing EDT (blue) and RT60(red) fitted to the decay curves.

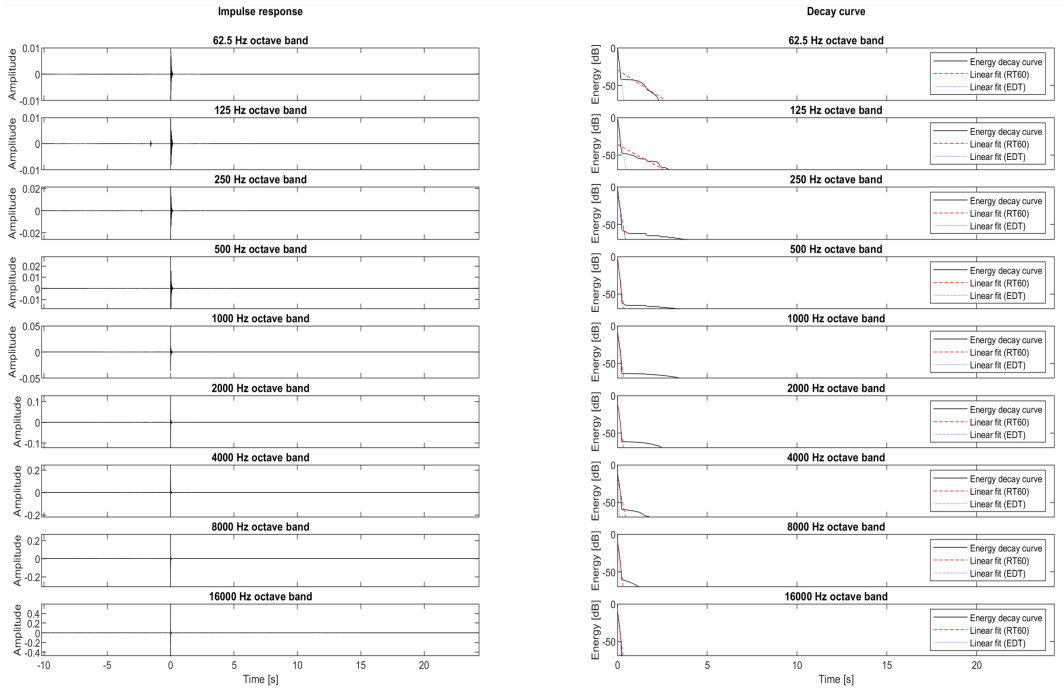

FIGURE A.2: Usability Lab (UL) RIR visualisation and energy decay curves over different octave bands, showing EDT (blue) and RT60(red) fitted to the decay curves.

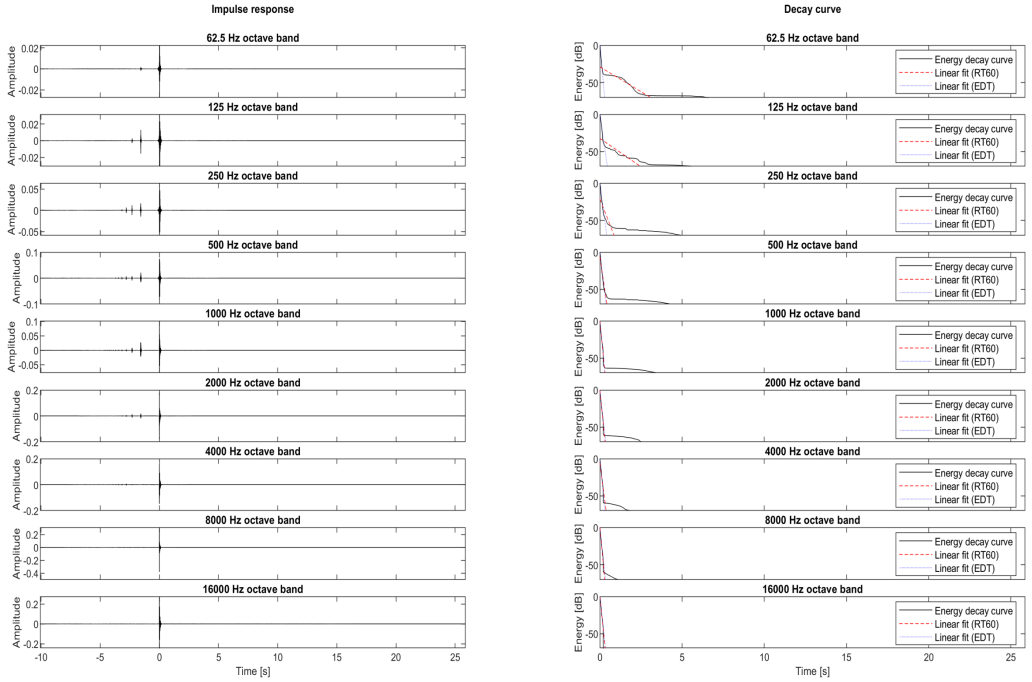

FIGURE A.3: Kitchen (KT) RIR visualisation and energy decay curves over different octave bands, showing EDT (blue) and RT60(red) fitted to the decay curves.

References

Ade20k dataset. https://tinyurl.com/ADE20K. [Online; accessed 2023-01-17].

Torii Akihiko, Imiya Atsushi, and Naoya Ohnishi. Two-and three-view geometry for spherical cameras. In *Workshop on omnidirectional vision, camera networks and non-classical cameras*, volume 105, pages 29–34, 2005.

Davide Anguita, Sandro Ridella, and Fabio Rivieccio. K-fold generalization capability assessment for support vector classifiers. In *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.*, pages 855–858, 2005.

Çamci Anil. 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, 2024.

Jiayang Ao, Qiuhong Ke, and Krista A Ehinger. Image amodal completion: A survey. *Computer Vision and Image Understanding*, 229:103661, 2023.

Iro Armeni, Sasha Sax, Amir R Zamir, and Silvio Savarese. Joint 2d-3d-semantic data for indoor scene understanding. *arXiv preprint arXiv:1702.01105*, 2017.

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

Mihai-Vlad Baran, Richard King, and Wieslaw Woszczyk. A general overview of methods for generating room impulse responses. *The Journal of the Acoustical Society of America*, 155(3_Supplement):A282–A282, 2024.

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

Matthew Berger, Andrea Tagliasacchi, Lee M Seversky, Pierre Alliez, Gael Guennebaud, Joshua A Levine, Andrei Sharf, and Claudio T Silva. A survey of surface reconstruction from point clouds. In *Computer graphics forum*, volume 36, pages 301–329. Wiley Online Library, 2017.

Mehmet Ilker Berkman. History of virtual reality. In *Encyclopedia of computer graphics and games*, pages 873–881. Springer, 2024.

Aurora Berni and Yuri Borgianni. Applications of virtual reality in engineering and product design: Why, what, how, when and where. *Electronics*, 9(7):1064, 2020.

Achintya K Bhowmik. Sensification of computing: adding natural sensing and perception capabilities to machines. *APSIPA Transactions on Signal and Information Processing*, 6:e1, 2017.

Avrim Blum, Adam Kalai, and John Langford. Beating the hold-out: Bounds for k-fold and progressive cross-validation. In *Proceedings of the twelfth annual conference on Computational learning theory*, pages 203–208, 1999.

Nikolas Borrel-Jensen. Accelerated methods for computing acoustic sound fields in dynamic virtual environments with moving sources. 2023.

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

Mark F Bradshaw, Andrew D Parton, and Richard A Eagle. The interaction of binocular disparity and motion parallax in determining perceived depth and perceived size. *Perception*, 27(11):1317–1331, 1998.

María A Bretos, Sergio Ibáñez-Sánchez, and Carlos Orús. Applying virtual reality and augmented reality to the tourism experience: a comparative literature review. *Spanish Journal of Marketing-ESIC*, 28(3):287–309, 2024.

Yingjie Cai, Xuesong Chen, Chao Zhang, Kwan-Yee Lin, Xiaogang Wang, and Hongsheng Li. Semantic scene completion via integrating instances and scene in-the-loop. In *CVPR*, pages 324–333, 2021.

Anh-Quan Cao and Raoul de Charette. Monoscene: Monocular 3d semantic scene completion. In *CVPR*, pages 3991–4001, 2022.

Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. *arXiv preprint arXiv:1709.06158*, 2017.

Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*, pages 801–818, 2018.

Mingfei Chen, Kun Su, and Eli Shlizerman. Be everywhere-hear everything (bee): Audio scene reconstruction by sparse audio-visual samples. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7853–7862, 2023.

Siyi Chen, Hermann J Müller, and Markus Conci. Amodal completion in visual working memory. *Journal of Experimental Psychology: Human Perception and Performance*, 42 (9):1344, 2016.

Weikai Chen, Cheng Lin, Weiyang Li, and Bo Yang. 3psdf: Three-pole signed distance function for learning surfaces with arbitrary topologies. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18522–18531, 2022.

Xiaokang Chen, Kwan-Yee Lin, Chen Qian, Gang Zeng, and Hongsheng Li. 3d sketch-aware semantic scene completion via semi-supervised structure prior. In *CVPR*, pages 4193–4202, 2020.

Yueh-Tung Chen, Martin Garbade, and Juergen Gall. 3d semantic scene completion from a single depth image using adversarial training. In *ICIP*, pages 1835–1839, 2019.

Corey I Cheng and Gregory H Wakefield. Introduction to head-related transfer functions (hrtfs): Representations of hrtfs in time, frequency, and space. In *Audio Engineering Society Convention 107*. Audio Engineering Society, 1999.

Ran Cheng, Christopher Agia, Yuan Ren, Xinhai Li, and Liu Bingbing. S3cnet: A sparse semantic scene completion network for lidar point clouds. In *Conference on Robot Learning*, pages 2148–2161. PMLR, 2021.

Mark Ming-Tso Chiang and Boris Mirkin. Intelligent choice of the number of clusters in k-means clustering: an experimental study with different cluster spreads. *Journal of classification*, 27:3–40, 2010.

Agata Ciekanowska, Adam Kiszczak-Gliński, and Krzysztof Dziedzic. Vr space such as. *Journal of Computer Sciences Institute*, 20:247–253, 2021.

Benjamin Coors, Alexandru Paul Condurache, and Andreas Geiger. Spherenet: Learning spherical representations for detection and classification in omnidirectional images. In *Proceedings of the European conference on computer vision (ECCV)*, pages 518–533, 2018.

Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *CVPR*, pages 9268–9277, 2019.

Thiago LT da Silveira, Paulo GL Pinto, Jeffri Murrugarra-Llerena, and Cláudio R Jung. 3d scene geometry estimation from 360 imagery: A survey. *ACM Computing Surveys*, 55(4):1–39, 2022.

Manuel Dahnert, Ji Hou, Matthias Nießner, and Angela Dai. Panoptic 3d scene reconstruction from a single rgb image. *Advances in Neural Information Processing Systems*, 34:8282–8293, 2021.

Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *CVPR*, pages 5828–5839, 2017a.

Angela Dai, Charles Ruizhongtai Qi, and Matthias Nießner. Shape completion using 3d-encoder-predictor cnns and shape synthesis. In *CVPR*, pages 5868–5877, 2017b.

Pei Dang, Jun Zhu, Yuxuan Zhou, Yuting Rao, Jigang You, Jianlin Wu, Mengting Zhang, and Weilian Li. A 3d-panoramic fusion flood enhanced visualization method for vr. *Environmental Modelling & Software*, 169:105810, 2023.

HEM Den Ouden, R Van Ee, and EHF De Haan. Colour helps to solve the binocular matching problem. *The Journal of Physiology*, 567(2):665–671, 2005.

Sanika Doolani, Callen Wessels, Varun Kanal, Christos Sevastopoulos, Ashish Jaiswal, Harish Nambiappan, and Fillia Makedon. A review of extended reality (xr) technologies for manufacturing training. *Technologies*, 8(4):77, 2020.

Aloisio Dourado, Teofilo E De Campos, Hansung Kim, and Adrian Hilton. Edgenet: Semantic scene completion from a single rgb-d image. In *ICPR*, pages 503–510, 2021.

Aloisio Dourado, Frederico Guth, and Teofilo de Campos. Data augmented 3d semantic scene completion with 2d segmentation priors. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 3781–3790, 2022.

Aloisio Dourado Neto. Towards complete 3d indoor scene understanding from a single point-of-view. 2024.

Dicle Dövencioğlu, Hiroshi Ban, Andrew J Schofield, and Andrew E Welchman. Perceptual integration for qualitatively different 3-d cues in the human brain. *Journal of Cognitive Neuroscience*, 25(9):1527–1541, 2013.

F Dunn, WM Hartmann, DM Campbell, and Neville H Fletcher. *Springer handbook of acoustics*. Springer, 2015.

David Eigen and Rob Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *Proceedings of the IEEE international conference on computer vision*, pages 2650–2658, 2015.

Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, pages 226–231, 1996.

Mark Everingham, SM Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *International journal of computer vision*, 111(1):98–136, 2015.

Hao Fang, Cihui Pan, and Hui Huang. Structure-aware indoor scene reconstruction via two levels of abstraction. *ISPRS Journal of Photogrammetry and Remote Sensing*, 178:155–170, 2021.

Angelo Farina. Simultaneous measurement of impulse response and distortion with a swept-sine technique. In *Audio engineering society convention 108*. Audio Engineering Society, 2000.

Angelo Farina. Advancements in impulse response measurements by sine sweeps. In *Audio engineering society convention 122*. Audio Engineering Society, 2007.

Shehzana Fatima. Accelerated inverse rendering using signed distance functions and ray-based implicit geometry models. 2024.

Hasan Baran Fırat, Luigi Maffei, and Massimiliano Masullo. 3d sound spatialization with game engines: the virtual acoustics performance of a game engine and a middleware for interactive audio design. *Virtual Reality*, 26(2):539–558, 2022.

Michael Firman, Oisin Mac Aodha, Simon Julier, and Gabriel J Brostow. Structured prediction of unobserved voxels from a single depth image. In *CVPR*, pages 5431–5440, 2016.

Ruochong Fu, Hang Wu, Mengxiang Hao, and Yubin Miao. Semantic scene completion with point cloud representation and transformer-based feature fusion. In *2023 IEEE International Conference on Image Processing (ICIP)*, pages 3369–3373. IEEE, 2023.

Thomas Funkhouser, Nicolas Tsingos, Ingrid Carlbom, Gary Elko, Mohan Sondhi, and James West. Modeling sound reflection and diffraction in architectural environments with beam tracing. In *Forum Acusticum*, page 8, 2002.

Thomas Funkhouser, Nicolas Tsingos, Ingrid Carlbom, Gary Elko, Mohan Sondhi, James E West, Gopal Pingali, Patrick Min, and Addy Ngan. A beam tracing method for interactive architectural acoustics. *The Journal of the acoustical society of America*, 115(2):739–756, 2004.

Shaohua Gao, Kailun Yang, Hao Shi, Kaiwei Wang, and Jian Bai. Review on panoramic imaging and its applications in scene understanding. *IEEE Transactions on Instrumentation and Measurement*, 71:1–34, 2022.

Martin Garbade, Yueh-Tung Chen, Johann Sawatzky, and Juergen Gall. Two stream 3d semantic scene completion. In *CVPRW*, pages 0–0, 2019.

Andreas Geiger and Chaohui Wang. Joint 3d object and layout inference from a single rgb-d image. In *German Conference on Pattern Recognition*, pages 183–195, 2015.

James Jerome Gibson. The senses considered as perceptual systems. 1966.

Ruiqi Guo, Chuhang Zou, and Derek Hoiem. Predicting complete 3d models of indoor scenes. *arXiv preprint arXiv:1504.02437*, 2015.

Yanwen Guo, Yuanqi Li, Dayong Ren, Xiaohong Zhang, Jiawei Li, Liang Pu, Changfeng Ma, Xiaoyu Zhan, Jie Guo, Mingqiang Wei, et al. Lidar-net: A real-scanned 3d point cloud dataset for indoor scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21989–21999, 2024.

Yu-Xiao Guo and Xin Tong. View-volume network for semantic scene completion from a single depth image. *arXiv preprint arXiv:1806.05361*, 2018.

Saurabh Gupta, Pablo Arbelaez, and Jitendra Malik. Perceptual organization and recognition of indoor scenes from rgb-d images. In *CVPR*, pages 564–571, 2013.

Saurabh Gupta, Ross Girshick, Pablo Arbeláez, and Jitendra Malik. Learning rich features from rgb-d images for object detection and segmentation. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VII 13*, pages 345–360. Springer, 2014.

Saurabh Gupta, Pablo Arbeláez, Ross Girshick, and Jitendra Malik. Aligning 3d models to rgb-d images of cluttered scenes. In *CVPR*, pages 4731–4740, 2015.

Daniel A Guttentag. Virtual reality: Applications and implications for tourism. *Tourism management*, 31(5):637–651, 2010.

E.A.P. Habets. *Single- and multi-microphone speech dereverberation using spectral enhancement*. Phd thesis 1 (research tu/e / graduation tu/e), Electrical Engineering, 2007.

Ankur Handa, Thomas Whelan, John McDonald, and Andrew J Davison. A benchmark for rgb-d visual odometry, 3d reconstruction and slam. In *2014 IEEE international conference on Robotics and automation (ICRA)*, pages 1524–1531. IEEE, 2014.

Ankur Handa, Viorica Patraucean, Vijay Badrinarayanan, Simon Stent, and Roberto Cipolla. Understanding real world indoor scenes with synthetic data. In *CVPR*, pages 4077–4085, 2016.

John A Hartigan and Manchek A Wong. Algorithm as 136: A k-means clustering algorithm. *Journal of the royal statistical society. series c (applied statistics)*, 28(1):100–108, 1979.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016a.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *ECCV*, pages 630–645, 2016b.

Yuwen Heng, Yihong Wu, Hansung Kim, and Srinandan Dasmahapatra. Cam-segnet: A context-aware dense material segmentation network for sparsely labelled datasets. In *17th International Conference on Computer Vision Theory and Applications (VISAPP)*, pages 190–201, 2022.

Yuwen Heng, Srinandan Dasmahapatra, and Hansung Kim. Dbat: Dynamic backward attention transformer for material segmentation with cross-resolution patches. *arXiv preprint arXiv:2305.03919*, 2023.

Takayuki Hidaka, Yoshinari Yamada, and Takehiko Nakagawa. A new definition of boundary point between early reflections and late reverberation in room impulse responses. *The Journal of the Acoustical Society of America*, 122(1):326–332, 2007.

Geoffrey Hinton. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Yaoshiang Ho and Samuel Wookey. The real-world-weight cross-entropy loss function: Modeling the costs of mislabeling. *IEEE access*, 8:4806–4813, 2019.

Difeng Hu, Vincent JL Gan, and Chao Yin. Robot-assisted mobile scanning for automated 3d reconstruction and point cloud semantic segmentation of building interiors. *Automation in Construction*, 152:104949, 2023.

Zeyu Hu, Xuyang Bai, Jiaxiang Shang, Runze Zhang, Jiayu Dong, Xin Wang, Guangyuan Sun, Hongbo Fu, and Chiew-Lan Tai. Voxel-mesh network for geodesic-aware 3d semantic segmentation of indoor scenes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

Binh-Son Hua, Quang-Hieu Pham, Duc Thanh Nguyen, Minh-Khoi Tran, Lap-Fai Yu, and Sai-Kit Yeung. Scenenn: A scene meshes dataset with annotations. In *2016 Fourth International Conference on 3D Vision (3DV)*, pages 92–101, 2016.

Binh-Son Hua, Minh-Khoi Tran, and Sai-Kit Yeung. Pointwise convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 984–993, 2018.

Jingwei Huang, Zhili Chen, Duygu Ceylan, and Hailin Jin. 6-dof vr videos with a single 360-camera. In *2017 IEEE Virtual Reality (VR)*, pages 37–44. IEEE, 2017.

MQ Huang, J Ninić, and QB Zhang. Bim, machine learning and computer vision techniques in underground construction: Current status and future perspectives. *Tunnelling and Underground Space Technology*, 108:103677, 2021.

Qiangui Huang, Weiyue Wang, and Ulrich Neumann. Recurrent slice networks for 3d segmentation of point clouds. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2626–2635, 2018.

Vedad Hulusic, Carlo Harvey, Kurt Debattista, Nicolas Tsingos, Steve Walker, David Howard, and Alan Chalmers. Acoustic rendering and auditory–visual cross-modal perception and interaction. In *Comput. Graph. Forum*, pages 102–131, 2012.

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

IoSR. Iosr matlab toolbox. https://github.com/IoSR-Surrey/MatlabToolbox/tree/master, 2024. Accessed: 2024-10-02.

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

Mingi Ji, Byeongho Heo, and Sungrae Park. Show, attend and distill: Knowledge distillation via attention-based feature matching. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 7945–7952, 2021.

Hao Jiang and Jianxiong Xiao. A linear approach to matching cuboids in rgbd images. In *CVPR*, pages 2171–2178, 2013.

Yue Jiang, Dantong Ji, Zhizhong Han, and Matthias Zwicker. Sdfdiff: Differentiable rendering of signed distance fields for 3d shape optimization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1251–1261, 2020.

Luis Guillermo Roldão Jimenez. *3D Scene Reconstruction and Completion for Autonomous Driving*. PhD thesis, Sorbonne Université, 2021.

Sam Kavanagh, Andrew Luxton-Reilly, Burkhard Wuensche, and Beryl Plimmer. A systematic review of virtual reality in education. *Themes in science and technology education*, 10(2):85–119, 2017.

Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? 30, 2017.

Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *CVPR*, pages 7482–7491, 2018.

Hansung Kim and Adrian Hilton. Block world reconstruction from spherical stereo image pairs. *Computer Vision and Image Understanding*, 139:104–121, 2015.

Hansung Kim, Luca Remaggi, Philip JB Jackson, and Adrian Hilton. Immersive spatial audio reproduction for vr/ar using room acoustic modelling from 360 images. In *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pages 120–126, 2019.

Hansung Kim, Luca Remaggi, Sam Fowler, Philip Jackson, and Adrian Hilton. Acoustic room modelling using 360 stereo cameras. *IEEE Transactions on Multimedia*, 2020a.

Hansung Kim, Luca Remaggi, Philip JB Jackson, and Adrian Hilton. Immersive virtual reality audio rendering adapted to the listener and the room. In *Real VR–Immersive Digital Reality: How to Import the Real World into Head-Mounted Immersive Displays*, pages 293–318. Springer, 2020b.

Hansung Kim, Luca Remaggi, Aloisio Dourado, Teofilo de Campos, Philip JB Jackson, and Adrian Hilton. Immersive audio-visual scene reproduction using semantic scene reconstruction from 360 cameras. *Virtual Reality*, pages 1–16, 2021.

Hansung Kim, Luca Remaggi, Aloisio Dourado, Teofilo de Campos, Philip JB Jackson, and Adrian Hilton. Immersive audio-visual scene reproduction using semantic scene reconstruction from 360 cameras. *Virtual Reality*, 26(3):823–838, 2022.

Josef Kittler, Mohamad Hatef, Robert PW Duin, and Jiri Matas. On combining classifiers. *IEEE transactions on pattern analysis and machine intelligence*, 20(3):226–239, 1998.

Homare Kon and Hideki Koike. Deep neural networks for cross-modal estimations of acoustic reverberation characteristics from two-dimensional images. In *Audio Engineering Society Convention 144*, 2018.

Bernd Krolla, Maximilian Diebold, Bastian Goldlücke, and Didier Stricker. Spherical light fields. In *BMVC*, number 67.1–67.12, 2014.

Hamid Laga, Laurent Valentin Jospin, Farid Boussaid, and Mohammed Bennamoun. A survey on deep learning techniques for stereo-based depth estimation. *IEEE transactions on pattern analysis and machine intelligence*, 44(4):1738–1764, 2020.

Scikit learn developers. Scikit-learn: Machine learning in python, 2007–2024. URL https://scikit-learn.org/stable/modules/clustering.html. © Copyright 2007–2024, Scikit-learn developers (BSD License).

Yeonkun Lee, Jaeseok Jeong, Jongseob Yun, Wonjune Cho, and Kuk-Jin Yoon. Spherephd: Applying cnns on a spherical polyhedron representation of 360deg images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9181–9189, 2019.

Tobias Lentz, Dirk Schröder, Michael Vorländer, and Ingo Assenmacher. Virtual reality system with integrated sound field simulation and reproduction. *EURASIP journal on advances in signal processing*, 2007:1–19, 2007.

Jie Li, Yu Liu, Dong Gong, Qinfeng Shi, Xia Yuan, Chunxia Zhao, and Ian Reid. Rgbd based dimensional decomposition residual network for 3d semantic scene completion. In *CVPR*, pages 7693–7702, 2019a.

Jie Li, Yu Liu, Xia Yuan, Chunxia Zhao, Roland Siegwart, Ian Reid, and Cesar Cadena. Depth based semantic scene completion with position importance aware loss. *IEEE Robotics and Automation Letters*, 5(1):219–226, 2019b.

Jie Li, Kai Han, Peng Wang, Yu Liu, and Xia Yuan. Anisotropic convolutional networks for 3d semantic scene completion. In *CVPR*, pages 3351–3359, 2020a.

Jie Li, Laiyan Ding, and Rui Huang. Imenet: Joint 3d semantic scene completion and 2d semantic segmentation through iterative mutual enhancement. In *IJCAI*, 2021.

Jie Li, Qi Song, Xiaohu Yan, Yongquan Chen, and Rui Huang. From front to rear: 3d semantic scene completion through planar convolution and attention-based network. *IEEE TMM*, 2023.

Shigang Li and K. Fukumori. Spherical stereo for the construction of immersive vr environment. In *IEEE Proceedings. VR 2005. Virtual Reality, 2005.*, pages 217–222, 2005. .

Siqi Li, Changqing Zou, Yipeng Li, Xibin Zhao, and Yue Gao. Attention-based multi-modal fusion network for semantic scene completion. In *AAAI*, pages 11402–11409, 2020b.

Song Li and Jürgen Peissig. Measurement of head-related transfer functions: A review. *Applied Sciences*, 10(14):5014, 2020.

Tong Li, Zhaoxuan Zhang, Yuxin Wang, Yan Cui, Yuqi Li, Dongsheng Zhou, Baocai Yin, and Xin Yang. Self-supervised indoor scene point cloud completion from a single panorama. *The Visual Computer*, pages 1–15, 2024.

Yunfeng Li and Zygmunt Pizlo. Depth cues versus the simplicity principle in 3d shape perception. *Topics in cognitive science*, 3(4):667–685, 2011.

Susan Liang, Chao Huang, Yapeng Tian, Anurag Kumar, and Chenliang Xu. Neural acoustic context field: Rendering realistic room impulse response with neural fields. *arXiv preprint arXiv:2309.15977*, 2023.

Yiyi Liao, Jun Xie, and Andreas Geiger. Kitti-360: A novel dataset and benchmarks for urban scene understanding in 2d and 3d. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):3292–3310, 2022.

Dahua Lin, Sanja Fidler, and Raquel Urtasun. Holistic scene understanding for 3d object detection with rgbd cameras. In *ICCV*, pages 1417–1424, 2013.

Di Lin, Haotian Dong, Enhui Ma, Lubo Wang, and Ping Li. Multi-head multi-scale feature fusion network for semantic scene completion. In *2023 International Conference on Artificial Intelligence and Education (ICAIE)*, pages 57–61. IEEE, 2023.

Shice Liu, Yu Hu, Yiming Zeng, Qiankun Tang, Beibei Jin, Yinhe Han, and Xiaowei Li. See and think: Disentangling semantic scene completion. 31, 2018.

Xianzhu Liu, Haozhe Xie, Shengping Zhang, Hongxun Yao, Rongrong Ji, Liqiang Nie, and Dacheng Tao. 2d semantic-guided semantic scene completion. *International Journal of Computer Vision*, pages 1–20, 2024.

Sagnik Majumder, Changan Chen, Ziad Al-Halah, and Kristen Grauman. Few-shot audio-visual learning of environment acoustics. *Advances in Neural Information Processing Systems*, 35:2522–2536, 2022.

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

Matteo Mascia, Antonio Canclini, Fabio Antonacci, Marco Tagliasacchi, Augusto Sarti, and Stefano Tubaro. Forensic and anti-forensic analysis of indoor/outdoor classifiers based on acoustic clues. In *2015 23rd European Signal Processing Conference (EUSIPCO)*, pages 2072–2076. IEEE, 2015.

Zihou Meng, Fengjie Zhao, and Mu He. The just noticeable difference of noise length and reverberation perception. In *2006 International Symposium on Communications and Information Technologies*, pages 418–421. IEEE, 2006.

Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4460–4470, 2019.

Henrik Møller, Michael Friis Sørensen, Dorte Hammershøi, and Clemen Boje Jensen. Head-related transfer functions of human subjects. *Journal of the Audio Engineering Society*, 43(5):300–321, 1995.

Brian CJ Moore. *An introduction to the psychology of hearing*. Brill, 2012.

Reza Moradi, Reza Berangi, and Behrouz Minaei. A survey of regularization strategies for deep models. *Artificial Intelligence Review*, 53(6):3947–3986, 2020.

Matej Močnik. pyrirtool: A python tool for room impulse response (rir) processing. `https://github.com/maj4e/pyrirtool`, 2023. Accessed: 2024-10-02.

Richard A Newcombe, Shahram Izadi, Otmar Hilliges, David Molyneaux, David Kim, Andrew J Davison, Pushmeet Kohi, Jamie Shotton, Steve Hodges, and Andrew Fitzgibbon. Kinectfusion: Real-time dense surface mapping and tracking. In *2011 10th IEEE international symposium on mixed and augmented reality*, pages 127–136. Ieee, 2011.

Yinyu Nie, Xiaoguang Han, Shihui Guo, Yujian Zheng, Jian Chang, and Jian Jun Zhang. Total3dunderstanding: Joint layout, object pose and mesh reconstruction for indoor scenes from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 55–64, 2020.

Isaac Kofi Nti, Owusu Nyarko-Boateng, and Justice Aning. Performance of machine learning algorithms with different k values in k-fold cross-validation. *International Journal of Information Technology and Computer Science*, 13(6):61–71, 2021.

NVIDIA. Segformer b5 finetuned ade 640x640. http://tinyurl.com/segformerb5, 2024. Accessed: 2024-02-06.

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

Jeong Joon Park, Peter R. Florence, Julian Straub, Richard A. Newcombe, and S. Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 165–174, 2019. URL https://api.semanticscholar.org/CorpusID:58007025.

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

Axel Plinge, Sebatian J Schlecht, Oliver Thiergart, Thomas Robotham, Olli Rummukainen, and Emanuël AP Habets. Six-degrees-of-freedom binaural audio reproduction of first-order ambisonics with distance information. In *Audio Engineering Society Conference: 2018 AES International Conference on Audio for Virtual and Augmented Reality*. Audio Engineering Society, 2018.

Archontis Politis, Sakari Tervo, Tapio Lokki, and Ville Pulkki. Parametric multidirectional decomposition of microphone recordings for broadband high-order ambisonic encoding. In *Audio Engineering Society Convention 144*. Audio Engineering Society, 2018.

Samuel F Potter, Maria K Cameron, and Ramani Duraiswami. Numerical geometric acoustics: An eikonal-based approach for modeling sound propagation in 3d environments. *Journal of Computational Physics*, 486:112111, 2023.

Alessandro Giuseppe Privitera, Federico Fontana, and Michele Geronazzo. The role of audio in immersive storytelling: a systematic review in cultural heritage. *Multimedia Tools and Applications*, pages 1–39, 2024.

Ville Pulkki and Matti Karjalainen. *Communication acoustics: an introduction to speech, audio and psychoacoustics*. John Wiley & Sons, 2015.

Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 652–660, 2017.

Nikunj Raghuvanshi, John Snyder, Ravish Mehra, Ming Lin, and Naga Govindaraju. Precomputed wave simulation for real-time sound propagation of dynamic sources in complex scenes. In *ACM SIGGRAPH 2010 papers*, pages 1–11. 2010.

Yongming Rao, Wenliang Zhao, Guangyi Chen, Yansong Tang, Zheng Zhu, Guan Huang, Jie Zhou, and Jiwen Lu. Denseclip: Language-guided dense prediction with context-aware prompting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18082–18091, 2022.

Anton Ratnarajah, Sreyan Ghosh, Sonal Kumar, Purva Chiniya, and Dinesh Manocha. Av-rir: Audio-visual room impulse response estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27164–27175, 2024.

Stephan Reichelt, Ralf Häussler, Gerald Fütterer, and Norbert Leister. Depth cues in human visual perception and their realization in 3d displays. In *Three-Dimensional Imaging, Visualization, and Display 2010 and Display Technologies and Applications for Defense, Security, and Avionics IV*, volume 7690, pages 92–103. SpIE, 2010.

Luca Remaggi, Philip Jackson, and Philip Coleman. Estimation of room reflection parameters for a reverberant spatial audio object. In *Audio Engineering Society Convention 138*, 2015.

Joachim Rix, Stefan Haas, and José Teixeira. *Virtual prototyping: Virtual environments and the product design process*. Springer, 2016.

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

Agnieszka Roginska and Paul Geluso. *Immersive Sound: The Art and Science of Binaural and Multi-Channel Audio*. Routledge, 2018. URL https://proquest.safaribooksonline.com/9781317480105.

Luis Roldao, Raoul De Charette, and Anne Verroust-Blondet. 3d semantic scene completion: a survey. *IJCV*, pages 1–28, 2022.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241, 2015.

Thomas Rossing. *Springer handbook of acoustics*. Springer Science & Business Media, 2007.

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

Peter J Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987.

Atul Rungta, Sarah Rust, Nicolas Morales, Roberta Klatzky, Ming Lin, and Dinesh Manocha. Psychoacoustic characterization of propagation effects in virtual environments. *ACM Transactions on Applied Perception (TAP)*, 13(4):1–18, 2016.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *IJCV*, 115(3):211–252, 2015.

Toni P Saarela and Michael S Landy. Integration trumps selection in object recognition. *Current Biology*, 25(7):920–927, 2015.

Aqsa Sabir, Rahat Hussain, Akeem Pedro, Mehrtash Soltani, Dongmin Lee, Chansik Park, and Jae-Ho Pyeon. Synthetic data generation with unity 3d and unreal engine for construction hazard scenarios: A comparative analysis.

Aqsa Sabir, Rahat Hussain, Akeem Pedro, Mehrtash Soltani, Dongmin Lee, Chansik Park, and Jae-Ho Pyeon. Synthetic data generation with unity 3d and unreal engine for construction hazard scenarios: A comparative analysis. In *International conference on construction engineering and project management*, pages 1286–1288. Korea Institute of Construction Engineering and Management, 2024.

Lauri Savioja and U Peter Svensson. Overview of geometrical room acoustic modeling techniques. *The Journal of the Acoustical Society of America*, 138(2):708–730, 2015.

Ashutosh Saxena, Jamie Schulte, Andrew Y Ng, et al. Depth estimation using monocular and stereo cues. In *IJCAI*, volume 7, pages 2197–2203, 2007.

Gabriel Schwartz and Ko Nishino. Recognizing material properties from images. *IEEE transactions on pattern analysis and machine intelligence*, 42(8):1981–1995, 2019.

Louena Shtrepi. Investigation on the diffusive surface modeling detail in geometrical acoustics based simulations. *The Journal of the Acoustical Society of America*, 145(3): EL215–EL221, 2019.

Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, pages 746–760, 2012.

Nikhil Singh, Jeff Mentch, Jerry Ng, Matthew Beveridge, and Iddo Drori. Image2reverb: Cross-modal reverb impulse response synthesis. In *ICCV*, pages 286–295, 2021.

Saptarshi Sinha, Hiroki Ohashi, and Katsuyuki Nakamura. Class-wise difficulty-balanced loss for solving class-imbalance. In *Proceedings of the Asian conference on computer vision*, 2020.

Shuran Song and Jianxiong Xiao. Sliding shapes for 3d object detection in depth images. In *ECCV*, pages 634–651, 2014.

Shuran Song and Jianxiong Xiao. Deep sliding shapes for amodal 3d object detection in rgb-d images. In *CVPR*, pages 808–816, 2016.

Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *CVPR*, pages 567–576, 2015.

Shuran Song, Fisher Yu, Andy Zeng, Angel X Chang, Manolis Savva, and Thomas Funkhouser. Semantic scene completion from a single depth image. In *CVPR*, pages 1746–1754, 2017.

Yu Song, Yiquan Wu, and Yimian Dai. A new active contour remote sensing river image segmentation algorithm inspired from the cross entropy. *Digital Signal Processing*, 48: 322–332, 2016.

Florian Spiess, Luca Rossetto, and Heiko Schuldt. Exploring multimedia vector spaces with vitrivr-vr. In *International Conference on Multimedia Modeling*, pages 317–323. Springer, 2024.

Guy-Bart Stan, Jean-Jacques Embrechts, and Dominique Archambeau. Comparison of different impulse response measurement techniques. *Journal of the Audio engineering society*, 50(4):249–262, 2002.

G Christopher Stecker, Travis M Moore, Monica Folkerts, Dmitry Zotkin, and Ramani Duraiswami. 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, 2018.

Rebecca Stewart and Mark Sandler. Statistical measures of early reflections of room impulse responses. In *Proc. of the 10th int. conference on digital audio effects (DAFx-07), Bordeaux, France*, pages 59–62, 2007.

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

Birger Streckel and Reinhard Koch. Lens model selection for visual tracking. In *Joint Pattern Recognition Symposium*, pages 41–48. Springer, 2005.

Peter Sturm. Pinhole camera model. In *Computer Vision: A Reference Guide*, pages 983–986. Springer, 2021.

David Stutz and Andreas Geiger. Learning 3d shape completion under weak supervision. *International Journal of Computer Vision*, 128(5):1162–1181, 2020.

Carole H Sudre, Wenqi Li, Tom Vercauteren, Sebastien Ourselin, and M Jorge Cardoso. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: Third International Workshop, DLMIA 2017, and 7th International Workshop, ML-CDS 2017, ...*, pages 240–248, 2017.

Baochen Sun and Kate Saenko. From virtual to reality: Fast adaptation of virtual object detectors to real domains. In *BMVC*, volume 1, page 3, 2014.

Muhammad Ali Syakur, B Khusnul Khotimah, EMS Rochman, and Budi Dwi Satoto. Integration k-means clustering method and elbow method for identification of the best customer profile cluster. In *IOP conference series: materials science and engineering*, volume 336, page 012017. IOP Publishing, 2018.

Richard Szeliski. *Computer vision: algorithms and applications*. Springer Nature, 2022.

Igor Szöke, Miroslav Skácel, Ladislav Mošner, Jakub Paliesek, and Jan Černockỳ. Building and evaluation of a real room impulse response dataset. *IEEE Journal of Selected Topics in Signal Processing*, 13(4):863–876, 2019.

Jiaxiang Tang, Xiaokang Chen, Jingbo Wang, and Gang Zeng. Not all voxels are equal: Semantic scene completion from the point-voxel perspective. In *AAAI*, pages 2352–2360, 2022.

Micah Taylor, Anish Chandak, Qi Mo, Christian Lauterbach, Carl Schissler, and Dinesh Manocha. Guided multiview ray tracing for fast auralization. *IEEE transactions on visualization and computer graphics*, 18(11):1797–1810, 2012.

Lyne Tchapmi, Christopher Choy, Iro Armeni, JunYoung Gwak, and Silvio Savarese. Segcloud: Semantic segmentation of 3d point clouds. In *2017 international conference on 3D vision (3DV)*, pages 537–547. IEEE, 2017.

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

Sam Van Damme, Maria Torres Vega, and Filip De Turck. Human-centric quality management of immersive multimedia applications. In *2020 6th IEEE Conference on Network Softwarization (NetSoft)*, pages 57–64. IEEE, 2020.

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

Tuomas Vuorinen. Animated sequence rendering performance comparison: Unity and unreal engine. https://www.theseus.fi/bitstream/handle/10024/863786/Vuorinen_Tuomas.pdf?sequence=2&isAllowed=y.

Fengyun Wang, Dong Zhang, Hanwang Zhang, Jinhui Tang, and Qianru Sun. Semantic scene completion with cleaner self. In *CVPR*, pages 867–877, 2023.

Fengyun Wang, Qianru Sun, Dong Zhang, and Jinhui Tang. Unleashing network potentials for semantic scene completion. In *CVPR*, pages 10314–10323, 2024a.

Qian Wang and Min-Koo Kim. Applications of 3d point cloud data in the construction industry: A fifteen-year review from 2004 to 2018. *Advanced engineering informatics*, 39:306–319, 2019.

Rui Wang, Yang Zhang, and Bing Jia. 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)*, pages 345–349. IEEE, 2021.

Xuzhi Wang, Di Lin, and Liang Wan. Ffnet: Frequency fusion network for semantic scene completion. In *AAAI*, pages 2550–2557, 2022.

Xuzhi Wang, Wei Feng, and Liang Wan. Multi-modal fusion architecture search for camera-based semantic scene completion. *Expert Systems with Applications*, 243: 122885, 2024b.

Zhengren Wang. 3d representation methods: A survey. *arXiv preprint arXiv:2410.06475*, 2024.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

Silvan Weder, Johannes L. Schönberger, Marc Pollefeys, and Martin R. Oswald. Neuralfusion: Online depth fusion in latent space. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3161–3171, 2020. URL https://api.semanticscholar.org/CorpusID:227227699.

Mingyun Wen and Kyungeun Cho. Object-aware 3d scene reconstruction from single 2d images of indoor scenes. *Mathematics*, 11(2):403, 2023.

Mario Wolf, Pascalis Trentsios, Niklas Kubatzki, Christoph Urbanietz, and Gerald Enzner. Implementing continuous-azimuth binaural sound in unity 3d. In *2020 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, pages 384–389. IEEE, 2020.

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

Shun-Cheng Wu, Keisuke Tateno, Nassir Navab, and Federico Tombari. Scfusion: Real-time incremental scene reconstruction with semantic completion. In *2020 International Conference on 3D Vision (3DV)*, pages 801–810, 2020.

Yihong Wu, Yuwen Heng, Mahesan Niranjan, and Hansung Kim. Depth estimation from a single omnidirectional image using domain adaptation. In *European Conference on Visual Media Production*, pages 1–9, 2021.

Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3d shapenets: A deep representation for volumetric shapes. In *CVPR*, pages 1912–1920, 2015.

Zeng Xiangyang, Chen Ke'an, and Sun Jincai. On the accuracy of the ray-tracing algorithms based on various sound receiver models. *Applied Acoustics*, 64(4):433–441, 2003.

Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M. Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. pages 12077–12090, 2021.

Yufei Xu, Qiming Zhang, Jing Zhang, and Dacheng Tao. Vitae: Vision transformer advanced by exploring intrinsic inductive bias. *Advances in neural information processing systems*, 34:28522–28535, 2021.

Zongyi Xu, Xiaoshui Huang, Bo Yuan, Yangfu Wang, Qianni Zhang, Weisheng Li, and Xinbo Gao. Retrieval-and-alignment based large-scale indoor point cloud semantic segmentation. *Science China Information Sciences*, 67(4):142104, 2024.

Donghai Yang, Yifan Liu, Qingjiu Chen, Meng Chen, Shaodong Zhan, Nim-kwan Cheung, Ho-Yin Chan, Zhidong Wang, and Wen Jung Li. Development of the high angular resolution 360° lidar based on scanning mems mirror. *Scientific Reports*, 13(1): 1540, 2023.

Jiawei Yao, Chuming Li, Keqiang Sun, Yingjie Cai, Hao Li, Wanli Ouyang, and Hongsheng Li. Ndc-scene: Boost monocular 3d semantic scene completion in normalized device coordinates space. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9421–9431. IEEE Computer Society, 2023.

Yiqun Yao and Rada Mihalcea. Modality-specific learning rates for effective multimodal additive late-fusion. In *The Association for Computational Linguistics (ACL)*, pages 1824–1834, 2022.

Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. Scannet++: A high-fidelity dataset of 3d indoor scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12–22, 2023.

Wangyang Yu and W Bastiaan Kleijn. Room acoustical parameter estimation from room impulse responses using deep neural networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:436–447, 2020.

Wentao Yuan, Tejas Khot, David Held, Christoph Mertz, and Martial Hebert. Pcn: Point completion network. In *2018 International Conference on 3D Vision (3DV)*, pages 728–737, 2018.

Biao Zhang and Peter Wonka. Point cloud instance segmentation using probabilistic embeddings. In *CVPR*, pages 8883–8892, 2021.

Jiahui Zhang, Hao Zhao, Anbang Yao, Yurong Chen, Li Zhang, and Hongen Liao. Efficient semantic scene completion network with spatial group convolution. In *ECCV*, pages 733–749, 2018a.

Liang Zhang, Le Wang, Xiangdong Zhang, Peiyi Shen, Mohammed Bennamoun, Guangming Zhu, Syed Afaq Ali Shah, and Juan Song. Semantic scene completion with dense crf from a single depth image. *Neurocomputing*, 318:182–195, 2018b.

Pingping Zhang, Wei Liu, Yinjie Lei, Huchuan Lu, and Xiaoyun Yang. Cascaded context pyramid for full-resolution 3d semantic scene completion. In *ICCV*, pages 7801–7810, 2019.

Yifan Zhang, Bingyi Kang, Bryan Hooi, Shuicheng Yan, and Jiashi Feng. Deep long-tailed learning: A survey. *IEEE TPAMI*, 2023.

Zhengyou Zhang. Microsoft kinect sensor and its effect. *IEEE multimedia*, 19(2):4–10, 2012.

Zhengyou Zhang. Camera parameters (intrinsic, extrinsic). In *Computer Vision: A Reference Guide*, pages 135–140. Springer, 2021.

Zichao Zhang, Henri Rebecq, Christian Forster, and Davide Scaramuzza. Benefit of large field-of-view cameras for visual odometry. In *2016 IEEE International Conference on Robotics and Automation (ICRA)*, pages 801–808. IEEE, 2016.

Borui Zhao, Quan Cui, Renjie Song, Yiyu Qiu, and Jiajun Liang. Decoupled knowledge distillation. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 11953–11962, 2022.

Min Zhong and Gang Zeng. Semantic point completion network for 3d semantic scene completion. In *European Conference on Artificial Intelligence*, pages 2824–2831. IOS Press, 2020.

Linqi Zhou, Yilun Du, and Jiajun Wu. 3d shape generation and completion through point-voxel diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5826–5835, 2021.

Nikolaos Zioulis, Antonis Karakottas, Dimitrios Zarpalas, and Petros Daras. Omnidepth: Dense depth estimation for indoors spherical panoramas. In *ECCV*, pages 448–465, 2018.

Michael Zollhöfer, Patrick Stotko, Andreas Görlitz, Christian Theobalt, Matthias Nießner, Reinhard Klein, and Andreas Kolb. State of the art on 3d reconstruction with rgb-d cameras. In *Computer graphics forum*, volume 37, pages 625–652. Wiley Online Library, 2018.