MDBNet 360°: 3D Audio-Visual Indoor Scene Reconstruction and Completion from a Single 360° RGB-D Image

Mona Alawadh[1,2] Atiyeh Alinaghi[1] Mahesan Niranjan[1] Hansung Kim[1]
[1]University of Southampton [2]Imam Mohammad Ibn Saud Islamic University
{m.alawadh, a.alinaghi, mn, H.Kim}@soton.ac.uk

Abstract

We introduce an approach for constructing immersive virtual spaces by generating comprehensive 3D voxelized models that encompass both geometric and semantic scene representations from a single 360° RGB-D input. The proposed approach utilizes a deep convolutional neural network to perform semantic scene reconstruction and completion, allowing for the estimation of complete scene semantics and geometries. This enables optimized spatial audio adaptation, ensuring that the sound environment is accurately tailored to the reconstructed space. To assess the acoustic properties, we measure parameters such as early decay time (EDT) and reverberation time (RT60) using the exponential sine sweep method (ESS). We used Unity with the Steam Audio plug-in for conducting simulations in virtual space. The proposed framework demonstrates better virtual space reconstruction and immersive sound generation, advancing semantically rich and spatially accurate virtual environments. Code and rendered sounds available at GitHub: https://github.com/blindRevAcc/Repo360/.

1. Introduction

Both visual and synchronised spatial audio are essential for creating truly immersive environment experiences [17, 33, 42]. 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 extended reality spaces, particularly in 3D visualisation and human-machine interaction [7, 8, 35, 37, 41, 43], there are still few studies focusing on the construction of 3D annotated models from 360° inputs. For instance, [23] 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 [20, 40, 45, 46], 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 360° inputs.

From the audio perspective, studies such as [6, 24, 27, 34, 39] 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 [16, 18], have integrated both audio and visual aspects to create more realistic and immersive virtual experiences. The study in [16] employed SegNet [3] 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 [18] employed EdgeNet [9], 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 [16] simplified scene objects using block representations, while the work in [18] 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 360° inputs while also capturing the RIRs for spatial sound evaluation to achieve a more immersive environment experience. In this research, we extend the previous work in [2], to develop a comprehensive 3D framework that integrates 360° RGB-D input. The pro-

posed 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 quality of the rendered sound within the full 3D scene by evaluating room impulse response (RIR) acoustic parameters, such as early decay time (EDT) [4] and reverberation time (RT60) [36]. We compare our results with state-of-the-art (SOTA) approaches using CVSSP dataset [3].

The summary of our contributions are as follows:

- Extend the work in [2] and propose MDBNet SSC model with a dual-head and combined loss function to train the model simultaneously with both RGB and depth inputs.
- 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.
- Perform an acoustic analysis of the 3D virtual environments generated by MDBNet360 through the evaluation of RIRs acoustic parameters, such as EDT and RT60, and comparing the results with SOTA methods. The proposed method showing better 3D scene reconstruction and acoustic parameters for the virtual space compared to SOTA.

2. Methodology

2.1. 3D SSC with MDBNet on Perspective Inputs

The architecture of the proposed MDBNet is depicted in Figure 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 flipped-Truncated Signed Distance Function (F-TSDF) for data representation within 3D space, which captures geometric information with dimensions of 240×144×240. We leverage the Segformer 'B5' model [48], a pre-trained transformer model for image semantic

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

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

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

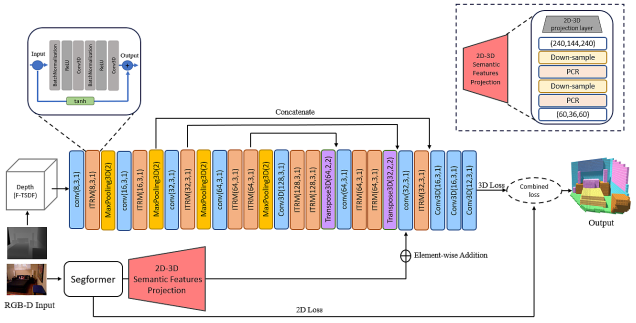

Figure 1. MDBNet is a dual-head network that processes 2D RGB semantics via a pre-trained Segformer with 2D-3D projection and geometric data via a 3D CNN with ITRM blocks. The network optimises a combined loss, which is a weighted sum of 3D loss and 2D semantics loss.

segmentation, to extract the 2D semantic features, which are subsequently projected into 3D space. Aligned with the projection method described in [25], 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 1:

$$p_{u,v} = K[R|t]p_{x,y,z}, \tag{1}$$

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 using late fusion since it provides the best results among early and middle fusion strategies. For late fusion we downsample the projected features using the Planar Convolution Residual (PCR) block [22], a variant of the Dimensional Decomposition Residual (DDR) block [20]. For the 3D input, we adopt the foundational structure of the 3D U-Net CNN, as utilised in [2], with a custom adaptation of the residual block. This adaptation includes implementing Identity Transformed within full pre-activation Residual Module (ITRM) by adding 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 within a normalized range, aligning with the inherent data range of TSDF or SDF, as demonstrated in [32, 47]. In the domain of SSC, the Tanh activation function has been applied to part of identity features, albeit in a different context [50]. Our research extends this exploration by investigating additional context for the application of Tanh. The model generates an output with a four-dimensional struc-

ture sized 60×36×60×12. The 12 channels represent the dataset classes ranging from 0 to 11. Class 0 is designated for empty spaces, whereas the remaining classes represent various object categories found in the NYUv2 [38] and NYUCAD [12] datasets, including ceiling, floor, wall, window, chair, bed, sofa, table, TV, furniture, and objects. 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 2:

$$L = \lambda L_{SS} + L_{SSC}. \tag{2}$$

Aligned with [45], 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 [2], evaluates the model's performance in 3D space, specifically using F-TSDF after integrating projected 2D semantic features in the current context. L_{SSC} employs a smoothed weights through an unsupervised clustering algorithm, K-means.

2.2. Extension to MDBNet360.

We extend MDBNet's inference capabilities to 360° RGB-D data by incorporating spherical-to-cubic projection and 3D transformation for comprehensive 3D reconstruction with 360° surroundings. The proposed design generates cubic views from 360° RGB-D input by converting the spherical RGB data into six perspective images. Following[15], 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 [18]. The Cartesian coordinates (x, y, z) are calculated using the latitude and longitude from the equirectangular image.

An occupancy voxel grid is constructed to represent the scene's surface. This is achieved by simulating four perspective views (left, front, right, and back) with each view rotated 90° around the Y-axis. The transformation of the Cartesian coordinates (x, y, z) for each view is performed using the following rotation matrix:

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

The F-TSDF is then calculated for each 3D view. The TSDF value represents the Euclidean distance of each voxel to the nearest surface voxel using specific truncation threshold t to reduce both computational load and memory usage within the perspective cubic view. The TSDF is flipped to provide strong gradients on surface [40] :

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

The sign in Equation 4 provides information about whether the voxel is in front of or behind the object's surface. In the F-TSDF representation, voxels in visible or empty spaces above surfaces are assigned values ranging from 0 to 1, while those in occluded areas are assigned values from -1 to 0, resulting in steep gradients at object surfaces. Then we pass the RGB perspective view with corresponding F-TSDF inputs into the proposed model. We construct a comprehensive inference pipeline by combining predictions from multiple MDBNet inferences. Our proposed architecture generates four 3D volumes, with boundary overlaps occurring between adjacent 3D views. These views are merged within a single comprehensive view using the summation rule [19] as illustrated in Figure 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:

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

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

Post-processing is applied to all inferred 3D views, including fitting planes (walls, ceiling, and floor) in the room to enhance overall scene quality, ensuring a more coherent and visually realistic representation. The 3D room, with the aggregated views, is then exported to Unity with Steam Audio for object's material assignment and sound rendering.

2.3. 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 captured by playing ESS audio on the virtual sound source within the 3D scene. The rendered sound is captured by the virtual listener. To generate the ESS audio, we follow the approach proposed by Farina [10, 11, 30]:

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

The virtual sound source sweeps through the t samples 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 8 and 9, respectively. The sweep has a duration of T.

$$\omega_1 = 2 \cdot \pi \cdot f_1/fs \quad (8)$$

$$\omega_2 = 2 \cdot \pi \cdot f_2/fs \quad (9)$$

The RIR is extracted by the deconvolution process of the recorded ESS, and then 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 [13]. 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 [14, 36]. 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 [4, 14]. 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 [16, 18]. 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 [28] and 5% for EDT [44].

3. Implementation and Experimental Setup

3.1. MDBNet

Training and Validation. We conduct our experiments using the PyTorch framework, on a single Nvidia RTX 8000 GPU. Both 2D and 3D network branches are trained simultaneously with MDBNet. Due to the two types of input representation we employ different learning rates to achieve effective performance as demonstrated in [49]. For the 2D input modality (RGB), we employ a pre-trained Segformer model, which is fine-tuned on the ADE20K dataset [1] at an image resolution of 640×640. The model weights are downloaded from Hugging Face [31]. 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 [45], we used the AdamW optimizer with 0.05 weight decay, and learning

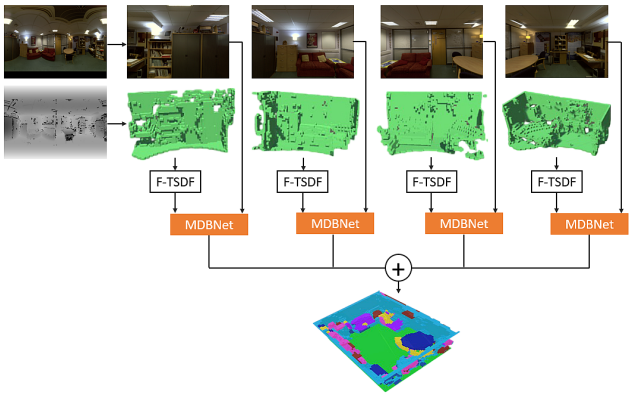

Figure 2. MDBNet360: RGB-D projection and prediction on full panorama MR scene from CVSSP dataset using MDBNet SSC model.

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 [29] 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.

Evaluation. 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 [40], 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

Table 1. 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

annotations and avoids misalignment.

Metrics We adopt Precision, Recall, and IoU as the evaluation measures for the SSC, following the approach of Song et al. [40]. 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. In this research, we follow [25] by evaluating all occluded occupied voxels and re-sampling empty occluded ones. As highlighted in [21, 26], 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.

Ablation Study. To confirm the impact of each component within our MDBNet, we modify the archetichture in [2] by integrating new components and conduct comprehensive experiments to evaluate their contributions, as detailed in Table 1. Initially, we train our model with RGB-D input and apply our combined loss defined in Section 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 [40]. 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%.

3.2. 3D Scenes Production Using MDBNet360.

We test the proposed method using the CVSSP dataset. 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 [18]. 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 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 (version: 2022.3.35f1), which is integrated with the Steam Audio plug-in (version: 4.5.3) for immersive sound rendering. Figure 2 illustrates the MR scene with 360° RGB-D spherical input, demonstrating scene partitions using our proposed method described in Section 2, accompanied by a comprehensive SSC 3D model prediction.

Table 2. 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

3.3. 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 2 lists the objects and their corresponding materials [18]. 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 2.3. 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. To generate spatialize sound, we choose the spatialize option and set the Spatial Blend to the 3D to generate immersive rendered sound. For the Steam Audio Source we apply HRTF-based binaural rendering, utilising the default Nearest interpolation option to control how HRTFs are adjusted as the sound source moves relative to the listener. The impact of HRTF is more pronounced in scenarios 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 2.3.

4. Results Analysis

4.1. 3D SSC of Perspective Scenes

Experiments were 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 3 and 4. Unlike previous studies, which did not specify the performance uncertainty, we averaged our scores across three folds to more accurately represent generalization performance. We compare MDBNet with SOTA methods that utilize hybrid architectures, focusing on voxel-based semantic segmentation on the NYUv2 dataset, as shown in Table 3. 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$_{Segformer}$ [46] which employed Segformer pretrained model for 2D RGB features, and PCANet [22], respectively. This establishes MDBNet as the new benchmark in SOTA performance. The efficacy of MDBNet is validated on the NYUCAD dataset, as shown in Table 4. Our average performance is competitive with AMMNet$_{Segformer}$ [46], and surpasses it when considering the upper-bound results. On the other hand, we provide qualitative analysis to illustrate the effectiveness of MDBNet's components. Figure 3 showcases various scenarios within the NYUCAD dataset, comparing when our combined loss function uses weighting based on re-sampling [40] within the 3D loss, when it applies class re-weighting [2], and when employing re-weighting [2] 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 3 in (a), (c), and (d). Additionally, our final design MDBNet offers better recognition of chairs with various shapes in the same figure in (b), (c), and (d), and it ensures enhanced differentiation between tables and chairs, as evident in (b) and (c). MDBNet model effectively recognizes challenging classes like windows and TVs, showcasing its robustness and adaptability. Additional results are available on our GitHub and supplementary materials.

4.2. 3D SSC of 360° Scenes

The original MDBNet demonstrated superior results, significantly outperforming other SSC models, such as EdgeNet [9], and Cleaners [45]. 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 [18] an extension of EdgeNet. We can clearly observe that MDBNet360 outperforms EdgeNet360 in semantic scene completion across

Table 3. Results on the NYUv2 dataset include averages and standard deviations for Precision, Recall, IoU, and mIoU metrics. The '*' represents the view-volume architecture type.

Method	Input	Res.	Scene Completion (SC)			Semantic Scene Completion (SSC)											
			Prec.	Recall	IoU	ceil.	floor	wall	win.	chair	bed	sofa	table	tvs	furn.	objs	mIoU
AMMNet$_{Segformer}$[46]	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[45]	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)[5]	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*[22]	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
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 4. Results on the NYUCAD dataset include averages and standard deviations for Precision, Recall, IoU, and mIoU metrics. The '*' represents the view-volume architecture type.

Method	Input	Res.	Scene Completion (SC)			Semantic Scene Completion (SSC)											
			Prec.	Recall	IoU	ceil.	floor	wall	win.	chair	bed	sofa	table	tvs	furn.	objs	mIoU
AMMNet$_{Segformer}$[46]	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)[5]	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
PCANet*[22]	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

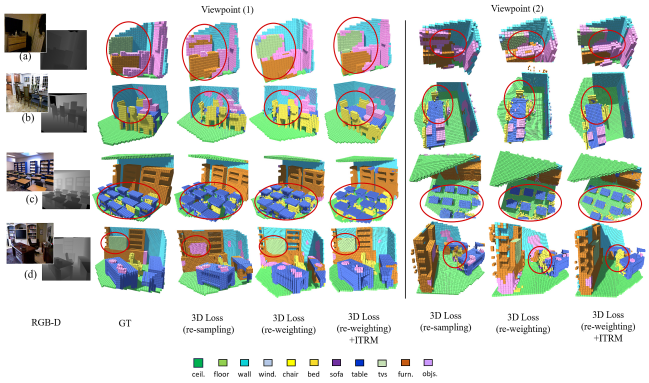

Figure 3. SSC results with different components on NYUCAD dataset. From left to right: (1) RGB-D input; (2) GT; (3) combined loss with re-sampling; (4) combined loss with re-weighting; (5) combined loss (using re-weighting) with ITRM blocks. Objects are color-coded, with circles highlighting key differences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

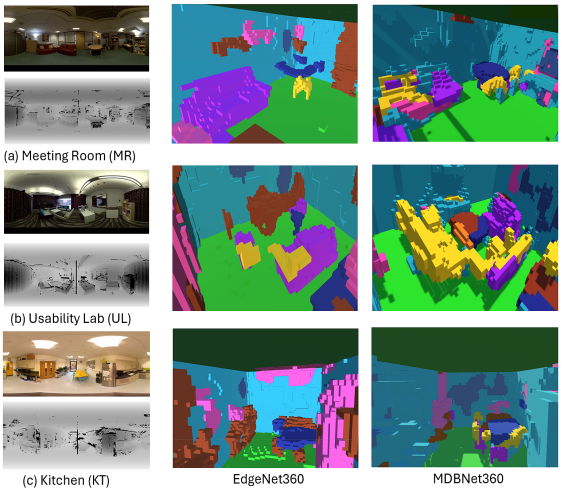

Figure 4. Qualitative comparison between MDBNet360 and EdgeNet360 on three scenes in CVSSP data. From top to bottom: MR, UL, and KT.

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

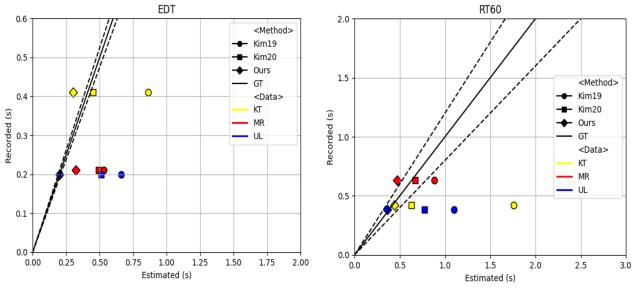

Figure 5. EDTs and RT60 for 3 CVSSP rooms related to the ground-truth (GT).

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

4.3. Spatial Audio within Virtual Space.

To provide a comprehensive evaluation of the Virtual space, we assess the sound quality within the virtual rooms generated by MDBNet360. Specifically, we evaluate the RIR based on the EDT and RT60 acoustic parameters. Our results are compared with the ground truth measurements obtained from sound modeled in real space, and SOTA models Kim19 [16] and Kim20 [18]. Overall, our approach demonstrates superior performance in both EDT and RT60 compared to Kim19 and Kim20, as shown in Figure 5. In the figure, the EDT scores for our model in the MR and UL scenes outperform those of Kim19 and Kim20, being closer to the ground truth. However, for the KT scene, the EDT score predicted by MDBNet360 is slightly shorter than the ground truth. We attribute this discrepancy to errors in the 3D semantics, where cabinets are mislabeled as 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. 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 [17]. This can be considered as a technical limitation of Steam Audio, the spatial audio rendering plug-in. This can be avoided by slight adjustment of the listener's position and fine-tuning of simulation parameters, such as the Reflection Mix Level, which helps to reduce the artifacts and provides more reliable results. Despite these challenges in the input context and variations across scenes, the final space reconstructed by MDBNet360 demonstrates better performance in both visual prediction and sound immersion compared to existing approaches. The rendered sound results are shared via Github account at: https://github.com/blindRevAcc/Repo360/.

5. Conclusion

In this work, we present a method for generating 3D virtual spaces that integrate both visual and acoustic cues from a single 360° RGB-D input. To this end, we develop MDBNet360, a model designed to produce a comprehensive 3D voxelized representation of indoor scenes. Our approach builds upon the pre-trained SSC MDBNet model, originally trained on the perspective-view NYUCAD dataset. We evaluate the acoustic quality of rendered sound using EDT and RT60 parameters. Our results show that the model effectively combines visual and acoustic cues, enhancing spatial audio realism and semantic representation in 3D scenes. This demonstrates the potential of our model to bridge the gap between visual fidelity and acoustic precision, providing a foundation for more immersive and interactive virtual environments using only a single 360° RGB-D input.

Acknowledgment

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

References

[1] Ade20k dataset. https://tinyurl.com/ADE20K. [Online; accessed 2023-01-17]. 4

[2] Mona Alawadh, Mahesan Niranjan, and Hansung Kim. 3d semantic scene completion from a depth map with unsuper-

vised learning for semantics prioritisation. In *2024 IEEE International Conference on Image Processing (ICIP)*, pages 3348–3354. IEEE, 2024. 1, 2, 3, 5, 6

[3] 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. 1

[4] Michael Barron. Interpretation of early decay times in concert auditoria. *Acta Acustica united with Acustica*, 81(4): 320–331, 1995. 2, 4

[5] 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 *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 324–333, 2021. 7

[6] 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. 1

[7] Agata Ciekanowska, Adam Kiszczak-Gliński, and Krzysztof Dziedzic. Vr space such as. *Journal of Computer Sciences Institute*, 20:247–253, 2021. 1

[8] 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. 1

[9] Aloisio Dourado, Teofilo E De Campos, Hansung Kim, and Adrian Hilton. Edgenet: Semantic scene completion from a single rgb-d image. In *Int. Conf. Pattern Recog.*, pages 503–510, 2021. 1, 6

[10] Angelo Farina. Simultaneous measurement of impulse response and distortion with a swept-sine technique. In *Audio engineering society convention 108*. Audio Engineering Society, 2000. 3

[11] Angelo Farina. Advancements in impulse response measurements by sine sweeps. In *Audio engineering society convention 122*. Audio Engineering Society, 2007. 3

[12] Michael Firman, Oisin Mac Aodha, Simon Julier, and Gabriel J Brostow. Structured prediction of unobserved voxels from a single depth image. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 5431–5440, 2016. 3

[13] 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. 4

[14] IoSR. Iosr matlab toolbox. https://github.com/IoSR-Surrey/MatlabToolbox/tree/master, 2024. Accessed: 2024-10-02. 4

[15] Hansung Kim and Adrian Hilton. Block world reconstruction from spherical stereo image pairs. *Computer Vision and Image Understanding*, 139:104–121, 2015. 3

[16] 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. 1, 4, 8

[17] 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, 2020. 1, 8

[18] 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. 1, 3, 4, 5, 6, 8

[19] 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. 3

[20] 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 *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 7693–7702, 2019. 1, 2

[21] Jie Li, Kai Han, Peng Wang, Yu Liu, and Xia Yuan. Anisotropic convolutional networks for 3d semantic scene completion. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3351–3359, 2020. 5

[22] 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 Trans. Multimedia*, 2023. 2, 6, 7

[23] 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. 1

[24] 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. 1

[25] Shice Liu, Yu Hu, Yiming Zeng, Qiankun Tang, Beibei Jin, Yinhe Han, and Xiaowei Li. See and think: Disentangling semantic scene completion. 31, 2018. 2, 5

[26] 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. 5

[27] 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. 1

[28] 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. 4

[29] Reza Moradi, Reza Berangi, and Behrouz Minaei. A survey of regularization strategies for deep models. *Artificial Intelligence Review*, 53(6):3947–3986, 2020. 4

[30] Matej Močnik. pyrirtool: A python tool for room impulse response (rir) processing. https://github.com/maj4e/pyrirtool, 2023. Accessed: 2024-10-02. 3

[31] NVIDIA. Segformer b5 finetuned ade 640x640. http://tinyurl.com/segformerb5, 2024. Accessed: 2024-02-06. 4

[32] 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. 2

[33] 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. 1

[34] 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. 1

[35] Joachim Rix, Stefan Haas, and José Teixeira. *Virtual prototyping: Virtual environments and the product design process*. Springer, 2016. 1

[36] 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. 2, 4

[37] 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. 1

[38] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *Eur. Conf. Comput. Vis.*, pages 746–760, 2012. 3

[39] Nikhil Singh, Jeff Mentch, Jerry Ng, Matthew Beveridge, and Iddo Drori. Image2reverb: Cross-modal reverb impulse response synthesis. In *Int. Conf. Comput. Vis.*, pages 286–295, 2021. 1

[40] Shuran Song, Fisher Yu, Andy Zeng, Angel X Chang, Manolis Savva, and Thomas Funkhouser. Semantic scene completion from a single depth image. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1746–1754, 2017. 1, 3, 4, 5, 6

[41] 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. 1

[42] 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. 1

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

[44] Michael Vorlander. International round robin on room acoustical computer simulations. *Proc. of 15ˆ¡ th¿ ICA, 1995. 6*, 1995. 4

[45] Fengyun Wang, Dong Zhang, Hanwang Zhang, Jinhui Tang, and Qianru Sun. Semantic scene completion with cleaner self. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 867–877, 2023. 1, 3, 4, 6, 7

[46] Fengyun Wang, Qianru Sun, Dong Zhang, and Jinhui Tang. Unleashing network potentials for semantic scene completion. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 10314–10323, 2024. 1, 6, 7

[47] 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. 2

[48] 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. 2

[49] 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. 4

[50] Pingping Zhang, Wei Liu, Yinjie Lei, Huchuan Lu, and Xiaoyun Yang. Cascaded context pyramid for full-resolution 3d semantic scene completion. In *Int. Conf. Comput. Vis.*, pages 7801–7810, 2019. 2