Room Acoustic Properties Estimation from a Single 360° Photo

Mona Alawadh
ECS, University of Southampton, UK
Imam Mohammad Ibn Saud Islamic University, Saudi Arabia
m.alawadh@soton.ac.uk

Yihong Wu
ECS, University of Southampton, UK
yihongwu@soton.ac.uk

Yuwen Heng
ECS, University of Southampton, UK
y.heng@soton.ac.uk

Luca Remaggi
Samsung R&D Institute, UK
lucaremaggi@gmail.com

Mahesan Niranjan
ECS, University of Southampton, UK
mn@ecs.soton.ac.uk

Hansung Kim
ECS, University of Southampton, UK
h.kim@soton.ac.uk

***Abstract*—Estimating room impulse responses (RIRs) in real spaces is a time-consuming and expensive process requiring multiple pieces of equipment, recordings, and processing. A simple computer-vision-based method from a single 360° photo is proposed to estimate the acoustic material properties of the space by reconstructing an approximated 3D geometry. A 3D semantic geometry model is reconstructed from a 360° image by monocular depth estimation and semantic scene completion. The material properties of semantic objects in the scene are estimated using the transformer-based dense material segmentation method. This model is used to simulate a 3D acoustic room model on the Unity platform with Steam spatial audio plug-in. Acoustic properties of the space are estimated from this virtual reproduction and evaluated against the actual ones in the real environment.**

***Index Terms*—3D reconstruction and completion, room acoustic modeling, depth estimation, material estimation.**

I. Introduction

Recently, research on combined audio-visual signal processing and rendering systems have been actively exploited as it gives better user experiences adapted to the human perceptual system [1]. The classical methods for estimating room acoustic properties are to use a complete audio system including microphones and loudspeakers, which requires time and resources [2]. This paper proposes a computer vision-based technique using a single 360° camera image to support room acoustic parameters estimation from room impulse responses (RIRs) for indoor scenes. Since a 3D geometry model with material properties can mimic real world acoustics [3], the proposed system can efficiently estimate room acoustic properties in a virtual space.

Many studies have been conducted for room geometry estimation and acoustic modeling. Some of them rely only on audio input [4], [5], while studies in [6]–[8] use visual input for spatial acoustic modelling. [9], [10] are based on combined audio-visual input. Our preliminary works tried to estimate room acoustics from a pair of stereo 360° images [11], [12]. However, these approaches require two synchronised and aligned 360° cameras to reconstruct 3D geometry, and material properties were manually assigned from object recognition results. In this paper, a simpler and more efficient system using only one 360° image is proposed by modifying and integrating our recent works on monocular depth estimation [13] and material estimation [14]. This single-camera approach can eliminate the restrictions of camera synchronisation and alignment, and can be easily extended to work on dynamic scenes. Material segmentation from visual input has been considered more difficult than object recognition, but recent work by [14]–[16] proved that combining contextual information (such as material boundary, object or scene labels) with material features extracted from image patches can boost the network performance in the material segmentation task.

In scene acoustics design and simulation, room impulse responses (RIRs) are analysed between a single audio source and a microphone to estimate room acoustics at the given location. The Unity virtual platform [17] combined with Steam Audio plug-in [18] is used for sound rendering and RIR measurement. Two room acoustic properties, early decay time (EDT) and reverberation time (RT60) are calculated from RIRs to evaluate the estimated room geometry and acoustics. The main contributions of this paper include:

- Present a complete pipeline room acoustic modeling using a 360° image.
- Monocular depth estimation algorithm for a 360° image.
- Materials estimation architecture for a 360° image.
- An evaluation of room acoustic modeling in a virtual space.

II. Proposed system

A. System Overview

This research aimed to develop an end-to-end system for 3D acoustic room modeling from a single 360° photo of an indoor scene to estimate EDT and RT60 in a virtual space. Figure 1 shows the flow of the proposed pipeline.

This work was supported by the EPSRC Programme Grant Immersive Audio-Visual 3D Scene Reproduction Using a Single 360 Camera (EP/V03538X/1).

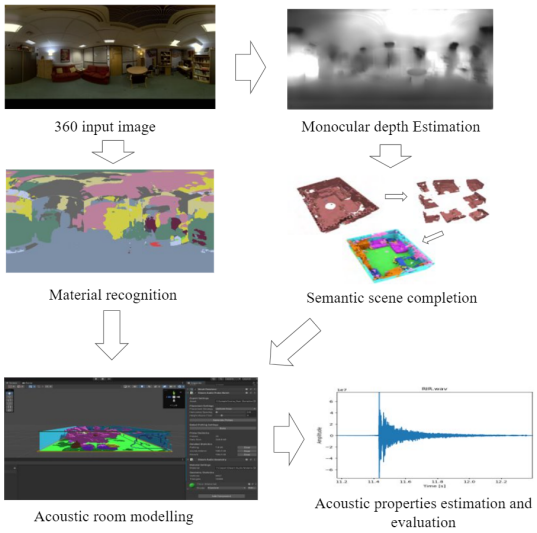

Fig. 1. End-to-end system structure: a single 360° image input to estimate monocular depth. Both materials recognition and 3D model reconstruction are processed in parallel. Results are integrated into Unity for complete 3D scene with materials labels to estimate and evaluate the acoustics environment

A complete indoor scene is captured using an off-the-shelf 360° camera. An omnidirectional monocular depth estimation using a supervised U-Net encoder-decoder architecture is applied to estimate the depth of the scene. A complete 3D geometrical structure is inferred by EdgeNet360 [19] which was designed for completing invisible parts of the 3D scene. On the other hand, material properties are estimated based on local and global features learning. The final acoustic room model is generated by integrating this information on the Unity virtual platform for sound reproduction and geometry rendering. By rendering sound with the 3D model in the virtual space, RIRs and other acoustic properties (EDT and RT60) are measured to evaluate the reproduced room model against the actual recorded sound in the space.

B. Monocular Depth Estimation

A U-Net shape encoder-decoder model for a single 360° image depth estimation is proposed by modifying our preliminary work based on domain adaptation [13]. We simplified the structure by removing the discriminator and focusing on supervised learning. It can achieve a higher accuracy of depth maps with more stable performance for predicting realistic scenes similar to the training dataset. For the encoder, ResNet50 [20] was used as the backbone, while the decoder consists of two convolution layers and four bi-linear up-sampling layers. Feature vectors extracted by the encoder are passed directly to the subsequent up-sampling layers in the decoder to infer corresponding depth maps. The training loss function (Equation 1) is a combination of two loss functions, including Structural Similarity (SSIM) [21] loss (Equation 2) and dense depth loss (Equation 3), whereas λ is a factor for dense depth loss and gt is the ground truth.

Fig. 2. Materials classes by the proposed material recognition module

$$L(gt, output) = \lambda L_{depth}(gt, output) + L_{SSIM}(gt, output) \tag{1}$$

$$L_{SSIM}(gt, output) = \frac{1 - SSIM(gt, output)}{2} \tag{2}$$

$$L_{depth}(gt, output) = \frac{1}{n}\sum_{p}^{n} |gt_p - output_p| \tag{3}$$

C. Materials Recognition

Inspired by the transformer architectures in [22] and [23] which take image patches as input and increase the receptive field by merging adjacent patches, our preliminary work on material estimation [14] has been modified by adopting windowed self-attention strategy [24] to control the patch size. In the proposed architecture, we can extract features from different patch sizes within a single network. The proposed network decides dependency on four patch sizes based on the input image rather than manually setting a fixed patch size for the whole dataset. In the implementation, the window size is set to two to learn features from image patches of sizes $\{8, 16, 32, 64\}$. To aggregate the features with the consideration of the input image, a set of attention masks (A_1, A_2, A_3, A_4) are predicted and normalised as in [25] from the final transformer layer. Finally, the merged feature is passed into the feature pyramid network [26] to recover the shape and predict the material labels for each pixel of the image. With this modification, the proposed method achieved an improvement of 15.12% on pixel accuracy with the local material database (LMD) compared with our preliminary work [14].

D. Semantic 3D Scene completion

Following our preliminary work in [12] and [19], a 3D voxel structure is reconstructed by projecting all points in the estimated depth maps into a 3D space. In order to cover the whole 360 surroundings, the 3D coordinate is partitioned into eight overlapped view parts from the scene center. The semantic scene completion using EdgeNet360 [27] is applied to individual areas and merged into one complete scene. The final inferred 3D model shows the scene reconstruction with semantic labels, and these labels are replaced by material labels inferred by the materials recognition module. Figure 2 shows the output classes of material recognition.

E. Sound Rendering and Room Acoustics Evaluation in a virtual Space

The reconstructed full 3D semantic scene is imported to the Unity platform with Steam Audio plug-in for room acoustics

simulation and sound rendering in a virtual space. Binaural sound is simulated between a virtual sound source and a listener with head related transfer functions (HRTFs) in the space. From this setting, binaural room impulse responses (BRIR) can be measured and analysed. The estimation and evaluation methods of room acoustic properties are inspired by [28]–[30]. BRIRs can be segmented into three parts: direct sound, early reflections, and late reverberations [4]. We analysed the EDT and RT60 of the generated sounds, as objective measures of their early reflections and late reverberation, respectively. EDT is a metric to evaluate the acoustics from adjacent reflectors by considering the energy carried by the early reflections. RT60 is related to the average absorption, location of room boundaries and size of the room, describing the reverberation from a physical point of view. EDT is calculated as six times the time required for the energy to decay 10 dB after the direct sound [31] and RT60 is measured as the time for the energy to decay 60 dB [32]. The average values over the 6-octave bands between 250 Hz and 8 kHz are reported for both EDT and RT60 in this research.

III. Experiments

This section shows the experimental results of the individual modules of the proposed pipeline and estimated acoustic properties of the scene. The proposed end-to-end system has been tested on two datasets: CVSSP [33] and 3D60 datasets [34]. The CVSSP set consists of five scenes with 360° image capture and ground-truth acoustic parameters measurement. We selected four scenes: Meeting Room (MR); Kitchen (KT); Usability Lab (UL); and Studio Hall (ST) for our experiments. The Listening Room (LR) in the CVSSP dataset was eliminated due to the use of acoustically controlled materials in this room. The 3D60 dataset is mainly used for omnidirectional depth estimation as it provides ground-truth depth data. All image sets, audio sources, and audio rendering results in this section are available at:
http://3dkim.com/research/VR/EUSIPCO.html

A. Omnidirectional Depth Estimation

Stanford2D3D and Matterport3D image sets with a resolution of 512×256 have been tested from the 3D60 dataset. They are office-room-based and house-based indoor scenes, respectively. All the scenes are 360° RGB images with their corresponding depth maps. As a pre-processing, scenes with over 5% of outliers are removed because some ground-truth depth maps in this dataset have large unknown/outlier areas. Stanford2D3D dataset ended up with 648 images for training and 82 images for testing, while Matterport3D contains 2075 images for training and 1144 images for testing.

Table I shows that the proposed model outperforms two state-of-the-art (SOTA) encoder-decoder models [34], [35] and get about 4% and 2% improvements in δ_1 accuracy, respectively. Similarly, the proposed method shows the best performance with the Matterport3D dataset as shown in Table II.

TABLE I
Performance comparison of depth estimation (Stanford2D3D)
(↑:the higher the better, ↓ the lower the better)

Model	$\delta_1 \uparrow$	$\delta_2 \uparrow$	$\delta_3 \uparrow$	$rel \downarrow$	$rms \downarrow$	$log_{10} \downarrow$
RectNet [34]	0.9102	0.9804	0.9902	0.0949	0.8573	0.0434
Alhashim and Wonka [35]	0.9323	0.9835	0.9906	0.0888	0.7956	0.0422
Proposed	**0.9519**	**0.9873**	**0.9918**	**0.0752**	**0.7894**	**0.0361**

TABLE II
Performance comparison of depth estimation (Matterport3D)

Model	$\delta_1 \uparrow$	$\delta_2 \uparrow$	$\delta_3 \uparrow$	$rel \downarrow$	$rms \downarrow$	$log_{10} \downarrow$
RectNet [34]	0.8885	0.9745	0.9909	0.1087	0.9355	0.0471
Alhashim and Wonka [35]	0.8996	0.9774	0.9918	0.1039	0.9017	0.0442
Proposed	**0.9055**	**0.9779**	**0.9919**	**0.0998**	**0.8755**	**0.0438**

Fig. 3. Depth estimation results (Left: RGB images, Middle: ground-truth depth maps, Right: Estimated depth maps)

Figure 3 shows the ground-truth data with white parts representing outliers caused by missing depth pixels. The proposed model accurately estimates the depth of input scenes and predicts missing parts in the depth.

B. Materials Estimation

Each 360° scene is projected into six partitions before the material recognition process and composed back to a 360° image with the resulting material labels. Table III illustrates the performance of the proposed materials estimation method compared with SOTA material segmentation models. The result reveals that the performance of the proposed method outperforms SOTA networks which were designed to extract information from full image rather than image patches. We also notice that apart from the antique ResNet, the remaining three SOTA models achieve comparable performance, despite the number of trainable parameters and the number of flops. This indicates that these networks reach the bottleneck of material segmentation task, by training with full images. Our transformer aggregates the features extracted from dynamic patches can break the bottleneck and improve the network generalisation ability, thus achieving a better performance.

C. 3D Acoustics Room Modeling

For acoustic properties evaluation, we followed the setting in our preliminary works in [11] and [12] and compared with them as there is no other similar approaches to the best of our knowledge. For simplicity in comparison, [11] will be referred to Model (1), and [12] Model (2) hereafter. The proposed system provides a full reconstruction from only one image while

TABLE III
PERFORMANCE COMPARISON OF MATERIAL RECOGNITION TRAINED ON LMD AGAINST SOTA SEGMENTATION NETWORKS

Metric	ResNet-152 [20]	ResNest-269e [36]	EfficientNet-b7 [37]	Swin-T [22]	**Proposed**
Pixel Acc	80.57	84.28	84.49	84.44	**86.77**
Mean Acc	74.12	79.55	78.17	78.60	**80.77**
# of param.	60.75	111	65.67	**29.52**	56.03
# of flops	70.27	128	35.30	**34.25**	41.23
FPS	31.35	11.92	18.87	**33.94**	27.44

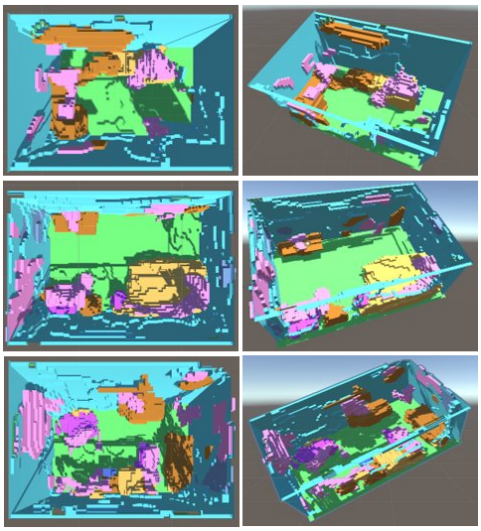

Fig. 4. 3D reconstructed models of the scenes in Figure 3 by the proposed method (Left: Top view, Right: Free viewpoint)

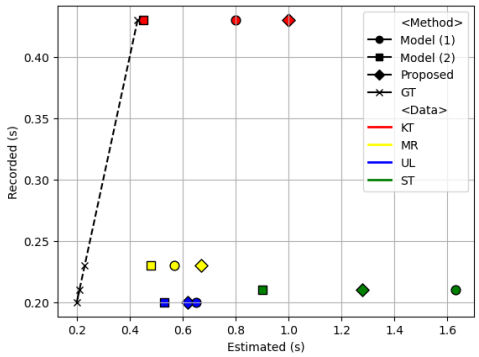

Fig. 5. EDTs for 4 CVSSP rooms related to the ground-truth (GT)

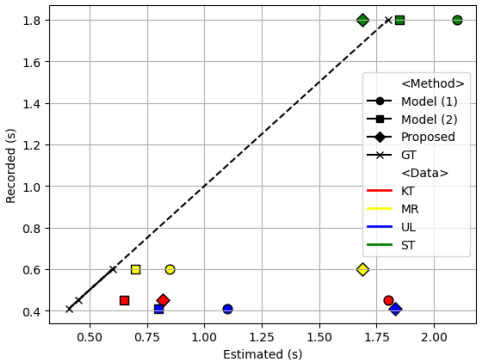

Fig. 6. RT60s for 4 CVSSP rooms related to the ground-truth (GT)

Model (1) and (2) use stereo image pairs. Therefore, we do not expect the proposed method outperforms the performance of Model (1) and (2), but demonstrate how close the proposed method can catch up with the reference methods with only one input image. Figure 4 illustrates snapshots of selected reconstructed models in Figure 3 by this pipeline.

In order to evaluate the estimated room geometry and acoustics, we simulated the same setting of the actual CVSSP data recordings in a virtual space with the reconstructed models, and compared the the virtual recordings with the actual ones. RIRs for the reproduced scenes in the virtual space were estimated by playing/recording a swept sine signal [38] with a virtual sound source and a virtual listener. Both virtual sound source and listener were located at the same positions where ground-truth data were recorded in the actual scenes. The recorded BRIRs are analysed to estimate EDT and RT60 as proposed in Section II-E. EDT is calculated as six times the energy decay 10 db by early reflections after direct sound [31]. While RT60 is more relates to room size and average of materials absorption in the room, considering the time required for energy to decay 60 dB [32]. Figure 5 and Figure 6 show the average results of EDT and RT60, respectively, over the 6 octave bands between 250 Hz and 8 kHz.

For both EDT and RT60 comparisons, the proposed single view method showed competitive performances with other stereo-based methods (Model (1) and (2)). The proposed method outperformed Model (1) in EDT with the UL and ST scenes owing to better estimation of objects and materials in the scenes. It also showed very good performance in RT60 with the KT and ST scenes. However, RT60 values for the MR and UL were too high. We found a scale issue in depth estimation in these cases. Actual MR and UL rooms are much smaller with low ceilings than most rooms in the training sets. Therefore, reconstructed scenes were larger than the ground-truth room volumes. This scene scale problem can be compensated by the scale factor when it's imported to Unity, but the scene/depth scale issue of the monocular depth estimation remains an open problem. Approximated 3D geometry also affected the reflection and reverberation properties of the room. Therefore, most of the simulated parameters show higher than ground-truth, but this happens even for Model (1) and (2). This can be compensated by tuning the acoustic parameters of materials, but we used the original material parameters in our experiments.

IV. CONCLUSION

In this work, an end-to-end acoustic room modelling system has been proposed to estimate room acoustic properties from a single 360° photo. From the input 360° image, real-scale depth field and scene material properties are estimated in parallel. A complete 3D geometry with material labels is constructed from these outputs. The 3D model is imported to the Unity platform with Steam Audio plug-in for virtual sound rendering. 3D spatial audio is rendered in the reproduced

virtual space by placing virtual sound source and listener. The reproduced room geometry and spatial audio were evaluated against actual data measured and recorded in the original rooms. The proposed method using only one image achieved competitive performance compared with the SOTA methods using aligned stereo image pair input. The proposed method can be applied more widely due to its simple set-up with only one camera. It can be extended to dynamic scene analysis with video streams as this system is free from camera alignment, calibration and synchronisation issues.

Future work will consider enhancing the depth scale for better 3D reconstruction. The material estimation part also requires further enhancement in matching the output material class types with actual acoustic parameters. Use of multi-modal audio-visual input for better geometry and acoustic property estimation will also be considered.

REFERENCES

[1] L. Remaggi, H. Kim, A. Neidhardt, A. Hilton, and P. J. Jackson, "Perceived quality and spatial impression of room reverberation in vr reproduction from measured images and acoustics," in *Proceedings of ICA*, 2019.

[2] H. Kon and H. Koike, "Deep neural networks for cross-modal estimations of acoustic reverberation characteristics from two-dimensional images," in *Audio Engineering Society Convention 144*. Audio Engineering Society, 2018.

[3] V. Hulusic, C. Harvey, K. Debattista, N. Tsingos, S. Walker, D. Howard, and A. Chalmers, "Acoustic rendering and auditory–visual cross-modal perception and interaction," in *Computer Graphics Forum*, vol. 31, no. 1. Wiley Online Library, 2012, pp. 102–131.

[4] W. Yu and W. B. Kleijn, "Room acoustical parameter estimation from room impulse responses using deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 436–447, 2020.

[5] D. D. Carlo, P. Tandeitnik, C. Foy, N. Bertin, A. Deleforge, and S. Gannot, "dechorate: a calibrated room impulse response dataset for echo-aware signal processing," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2021, no. 1, pp. 1–15, 2021.

[6] H. Kim, L. Remaggi, S. Fowler, P. Jackson, and A. Hilton, "Acoustic room modelling using 360 stereo cameras," *IEEE Transactions on Multimedia*, 2020.

[7] H. Kim, R. J. Hughes, L. Remaggi, P. J. Jackson, A. Hilton, T. J. Cox, and B. Shirley, "Acoustic room modelling using a spherical camera for reverberant spatial audio objects," in *Audio Engineering Society Convention 142*. Audio Engineering Society, 2017.

[8] R. Garg, R. Gao, and K. Grauman, "Geometry-aware multi-task learning for binaural audio generation from video," *arXiv preprint arXiv:2111.10882*, 2021.

[9] D. Li, T. R. Langlois, and C. Zheng, "Scene-aware audio for 360° videos," *ACM Trans. Graph.*, vol. 37, no. 4, jul 2018. [Online]. Available: https://doi.org/10.1145/3197517.3201391

[10] L. Remaggi, H. Kim, P. J. Jackson, and A. Hilton, "An audio-visual method for room boundary estimation and material recognition," in *Proceedings of the 2018 Workshop on Audio-Visual Scene Understanding for Immersive Multimedia*, 2018, pp. 3–9.

[11] H. Kim, L. Remaggi, P. J. Jackson, and A. 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)*. IEEE, 2019, pp. 120–126.

[12] H. Kim, L. Remaggi, A. Dourado, T. d. Campos, P. J. Jackson, and A. Hilton, "Immersive audio-visual scene reproduction using semantic scene reconstruction from 360 cameras," *Virtual Reality*, pp. 1–16, 2021.

[13] Y. Wu, Y. Heng, M. Niranjan, and H. Kim, "Depth estimation from a single omnidirectional image using domain adaptation," in *European Conference on Visual Media Production*, 2021, pp. 1–9.

[14] Y. Heng, Y. Wu, H. Kim, and S. 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)*, vol. 5, 2022, pp. 190–201.

[15] G. Schwartz and K. Nishino, "Material recognition from local appearance in global context," in *Biol. and Artificial Vision (Workshop held in conjunction with ECCV 2016)*, 2016.

[16] ——, "Recognizing material properties from images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 8, pp. 1981–1995, 2020.

[17] "Unity," 2022, https://unity.com/, Last accessed on 2022-02-21.

[18] "Steam audio," 2022, https://valvesoftware.github.io/steam-audio/, Last accessed on 2022-02-21.

[19] A. Dourado, H. Kim, T. E. de Campos, and A. Hilton, "Semantic scene completion from a single 360-degree image and depth map." in *VISIGRAPP (5: VISAPP)*, 2020, pp. 36–46.

[20] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[21] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[22] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10 012–10 022.

[23] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[24] S. Hofstätter, H. Zamani, B. Mitra, N. Craswell, and A. Hanbury, "Local self-attention over long text for efficient document retrieval," in *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2020, pp. 2021–2024.

[25] A. Tao, K. Sapra, and B. Catanzaro, "Hierarchical multi-scale attention for semantic segmentation," *arXiv preprint arXiv:2005.10821*, 2020.

[26] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2117–2125.

[27] A. Dourado, T. E. de Campos, H. Kim, and A. Hilton, "Edgenet: Semantic scene completion from rgb-d images," *arXiv preprint arXiv:1908.02893*, vol. 1, 2019.

[28] P. Coleman, A. Franck, P. J. Jackson, R. J. Hughes, L. Remaggi, and F. Melchior, "Object-based reverberation for spatial audio," *Journal of the Audio Engineering Society*, vol. 65, no. 1/2, pp. 66–77, 2017.

[29] L. Remaggi, P. Jackson, and P. Coleman, "Estimation of room reflection parameters for a reverberant spatial audio object," in *Audio Engineering Society Convention 138*. Audio Engineering Society, 2015.

[30] P. Coleman, A. Franck, D. Menzies, and P. J. Jackson, "Object-based reverberation encoding from first-order ambisonic rirs," in *Audio Engineering Society Convention 142*. Audio Engineering Society, 2017.

[31] M. Barron, "Interpretation of early decay times in concert auditoria," *Acta Acustica united with Acustica*, vol. 81, no. 4, pp. 320–331, 1995.

[32] F. Dunn, W. Hartmann, D. Campbell, and N. H. Fletcher, *Springer handbook of acoustics*. Springer, 2015.

[33] "Cvssp dataset," 2022, http://3dkim.com/research/VR/index.html, Last accessed on 2022-02-21.

[34] N. Zioulis, A. Karakottas, D. Zarpalas, and P. Daras, "Omnidepth: Dense depth estimation for indoors spherical panoramas," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 448–465.

[35] I. Alhashim and P. Wonka, "High quality monocular depth estimation via transfer learning," *arXiv preprint arXiv:1812.11941*, 2018.

[36] H. Zhang, C. Wu, Z. Zhang, Y. Zhu, H. Lin, Z. Zhang, Y. Sun, T. He, J. Mueller, R. Manmatha *et al.*, "Resnest: Split-attention networks," *arXiv preprint arXiv:2004.08955*, 2020.

[37] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International conference on machine learning*. PMLR, 2019, pp. 6105–6114.

[38] A. Farina, "Simultaneous measurement of impulse response and distortion with a swept-sine technique," in *Audio engineering society convention 108*. Audio Engineering Society, 2000.