[Appendix] Dense 2D-3D Indoor Prediction with Sound via Aligned Cross-Modal Distillation

A. Implementation Details

We use a 96×257 binaural audio spectrogram as an input for the student model, where we apply a short-time Fourier transform with a signal length of 512, hop length of 12, and window length of 40. We use Adam optimizer [1] with the learning rate of 0.0001 and polynomial scheduling of 0.9 for 60 epochs.

A.1. Depth Estimation

Following Jiang *et al.* [2], we clip the output values to [0.01, 10]. For the teacher model, we employ 192 × 384 panoramic images as input and apply random rotation (*i.e.*, roll along the horizontal axis) for data augmentation. Random audio channel flipping is also used as a data augmentation method during the student model training, corresponding to a horizontal image flip.

A.2. Semantic Segmentation

We use the same configuration as in the depth estimation task for input resolution and data augmentation. Instead of 40 labels in the Matterport3D [3], we employ nine categories for segmentation: wall, floor, chair, door, table, window, bed, ceiling, and column. We merge other classes that have similar semantics with the selected nine categories, e.g., picture→wall, sofa→chair, etc. The remaining categories are grouped into one miscellaneous category, which is not taken into consideration during training. Pixels classified as miscellaneous constitute 4.65% of all pixels in the training split. Since we utilize annotations from 3D meshes, we filter out noisy observations within the nine categories, discarding categories with less than 1% of all pixels for each image.

A.3. 3D Scene Reconstruction

Using a 16^3 voxel grid input, the teacher model predicts a 32^3 voxel grid. Each grid cell in the 32^3 voxel output has a value of 0 or 1, indicating whether the space is occupied or not. We compute the loss using pseudo-ground truth voxels and binary cross-entropy loss function, which helps bring the model close to the desired occupancy grid output. For evaluation, we employ the marching cubes algorithm [4] to produce a mesh from the voxel grid. Except for IoU, we calculate normal completeness, Chamfer distance, and F1 score using the generated mesh.

		MAE$_\downarrow$	RMSE$_\downarrow$	$\delta_{1\uparrow}$
$d \leq 1$	Pseudo-GT ($\mathcal{L}_p$) [5]	0.9539	1.6187	0.6378
	+ Rank [6]	0.9576	1.6184	0.6367
	+ **SAM$_{3,4}$**	**0.9193**	**1.5773**	**0.6659**
$d \leq 2$	Pseudo-GT ($\mathcal{L}_p$) [5]	0.9905	1.6502	0.6274
	+ Rank [6]	0.9962	1.6567	0.6241
	+ **SAM$_{3,4}$**	**0.9860**	**1.6477**	**0.6474**
$d \leq 3$	Pseudo-GT ($\mathcal{L}_p$) [5]	**1.0226**	1.6866	0.6179
	+ Rank [6]	1.0241	1.6870	0.6161
	+ **SAM$_{3,4}$**	1.0337	**1.6838**	**0.6248**
$d \leq 4$	Pseudo-GT ($\mathcal{L}_p$) [5]	1.0668	1.7338	0.6020
	+ Rank [6]	1.0638	1.7276	0.6034
	+ **SAM$_{3,4}$**	**1.0473**	**1.7013**	**0.6174**

Table 6: Comparison of non-identical emitter-receiver pairs on DAPS-Depth test split.

A.4. Computation Environment

1. GPU: NVIDIA RTX A6000
2. CPU: Intel(R) Xeon(R) Gold 6130 CPU
3. OS: Ubuntu 18.04 LTS
4. RAM: SAMSUNG DDR4 8G
5. Relevant software libraries: Anaconda distribution of python (3.7) and PyTorch (1.12)

Please refer to our source code for more details.

B. Additional Analysis

B.1. Influence of Auditory Observations

Table 6 reports the performance of different cross-modal distillation methods on DAPS-Depth test split under non-identical emitter-receiver pairs for audio inputs. To analyze the robustness against sound sources from varying locations, we include all pairs whose distance between an emitter and a receiver is up to four meters ($d \leq 4$). Since the appearance of the spectrogram may significantly vary with

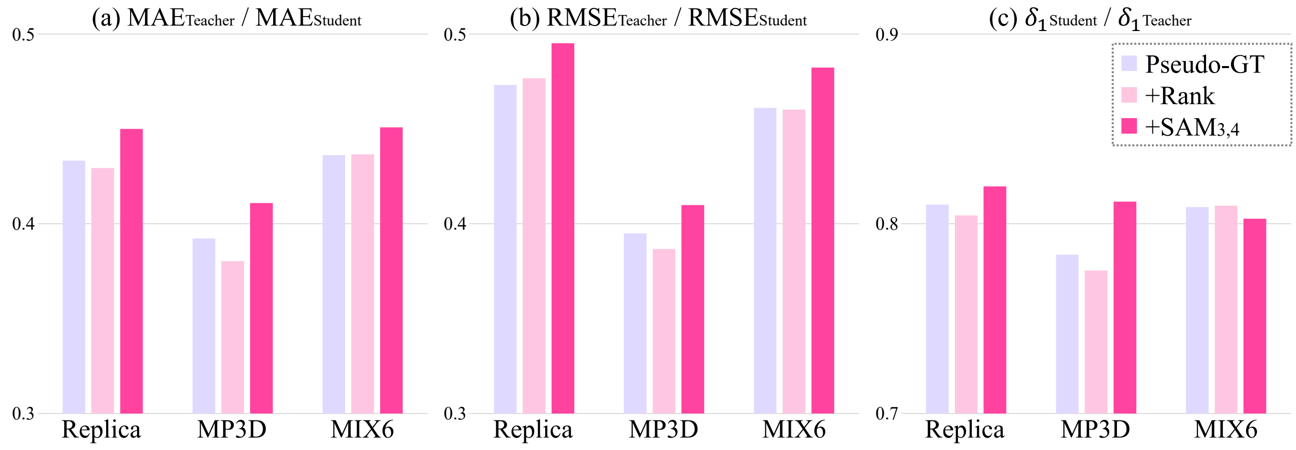

Figure 5: Influence of teacher model pre-training for audio-based depth-estimation on Replica. Higher is better.

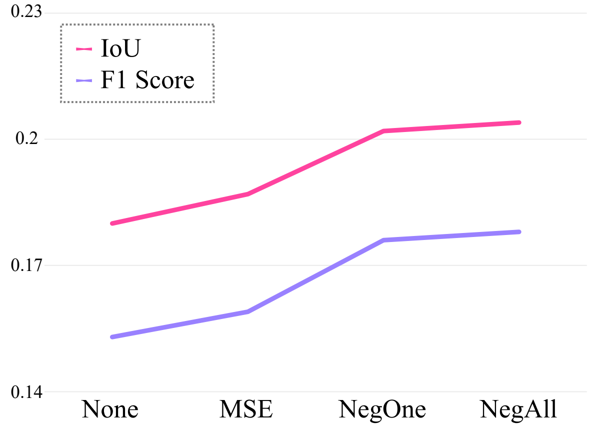

Figure 6: Influence of feature loss on DAPS-3D test split.

respect to the location of sound sources, the performance decays as the distance becomes greater. Still, our method displays superior performance for most of the configurations.

B.2. Influence of Learnable Spatial Embeddings

Table 7 reports the performance of our distillation framework with respect to the number of learnable spatial embeddings (K). To better analyze the influence of embeddings, we exclude the multi-head attention from the SAM block using U-Net as a backbone, *i.e.*, U-Net+SAM$_{\text{SpatialEmbeddings}}$. In general, more spatial embeddings employed throughout the spatial alignment via matching process implies better performance, where we select $K = 64$ by reflecting both performance and memory. A notable performance degradation in $K = 1$ indicates that it is not sufficient to leverage a single spatial embedding per position to reconstruct dense output.

K	MAE$_\downarrow$	RMSE$_\downarrow$	$\delta_{1\uparrow}$	$\delta_{2\uparrow}$	$\delta_{3\uparrow}$
1	1.3337	1.9798	0.4824	0.6656	0.7772
4	0.8760	1.5560	**0.6839**	0.8310	0.8949
16	0.8756	1.5450	0.6782	0.8311	0.8974
64	0.8739	1.5378	0.6738	**0.8313**	**0.9000**
256	**0.8700**	**1.5367**	0.6780	0.8306	0.8975

Table 7: Influence of the number of learnable spatial embeddings K on DAPS-Depth test split.

B.3. Influence of Cross-Dataset Transfer

Fig. 5 compares the performance of various methods on Replica [7] using different pre-trained teacher models. We follow Chen *et al*. [8] for the train, validation, and test split of Replica. We use three distinct pre-trained teacher models: the model trained on Replica, Matterport3D, and MIX6 [9], respectively. For the teacher model trained on MIX6, a large-scale dataset for depth estimation on a normal field-of-view image, we leverage pre-trained weight from Ranftl *et al*. [10] and fine-tune the last three convolutional layers with Replica to adjust the mean and variance of the prediction.

The error rates of different teacher models are similar to each other, *i.e.*, MAE of 0.2134, 0.2040, and 0.2144, respectively. The performance of Replica-trained teacher relatively falls short due to the small scale of the dataset, while MIX6-trained teacher mildly suffers from the domain gap between the datasets. In varying teacher models for cross-modal distillation, our approach consistently improves the relative performance regardless of the training dataset of teacher models. Fig. 10 visualizes the audio-based depth estimation results of our approach on Replica.

B.4. Extension of 3D Scene Reconstruction

We further demonstrate the effectiveness of our approach in non-iid settings for audio-based 3D scene reconstruction.

To be specific, we distill the knowledge from the teacher model trained on DAPS-3D (*i.e.*, Matterport3D) for Replica test split mentioned in Sec. B.3. Table 8 reports the reconstruction performance on Replica, where our method outperforms prior arts in all four metrics as in the iid setting (*i.e.*, Table 4).

Fig. 12 displays the qualitative results of audio-based 3D scene reconstruction on Replica test split. Due to the domain gap between the datasets, it is relatively more challenging to make precise predictions on Replica. Nonetheless, our approach, in general, is capable of identifying the size and structure of a surrounding scene. Additionally, it can recognize and generate the shape of objects, providing a subtle indication of their presence in the final output.

B.5. Influence of Feature Loss

Fig. 6 illustrates the model's performance with varying feature loss implementations using ConvONet+SAM$_{3,4}$ as a backbone. Our loose triplet-based learning objective significantly contributes to the performance, increasing IoU by 16% and F1 by 13%, respectively. On the other hand, forcefully matching the two heterogeneous feature maps by minimizing MSE has minimal impact on the performance. When training with our feature loss, using all the other features in a_i as negative samples (NegAll) is better than randomly choosing one of the neighboring features as a negative sample (NegOne).

B.6. More Qualitative Examples

Fig. 9 illustrates additional qualitative examples of audio-based semantic segmentation on DAPS-Semantic test split. Our method precisely predicts the room's layout while discerning semantic objects like a table with chairs or a doorframe. Fig. 11 visualizes more qualitative examples of audio-based 3D scene reconstruction on DAPS-3D test split.

B.7. Analysis on Real-world Audio Sample

To investigate the performance of our framework with real-world audio-based dense prediction, we apply our proposed framework to BatVision data [11] that addresses depth estimation in a restricted field of view. We observe that our approach can also be effective in real-world indoor/outdoor scenarios, as visualized in Fig. 7.

B.8. Analysis on limitation

To better understand the limitations of our approach, we categorize the failure modes in each task, as exemplified in Fig. 8. We analyze that there are two major causes behind inaccurate prediction: sophisticated furnishings and ambiguous layouts. For example, the first row of Fig. 8 involves a corridor filled with a variety of furniture and doors, making it hard to capture the surface of the room accurately. The second row displays a scene with an open, reverberant hall with pillars that considerably affects the room acoustics, which makes the model prone to mispredict the correct layout.

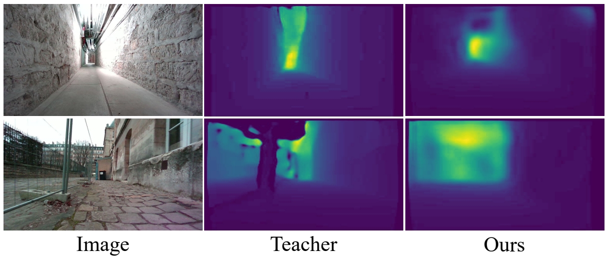

Figure 7: Qualitative examples on real-world samples.

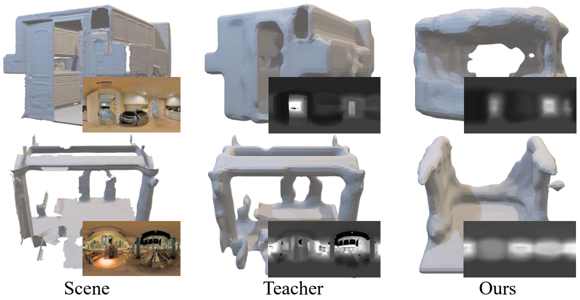

Figure 8: Failure mode analysis.

		IoU$_\uparrow$	Chamfer$_\downarrow$	NC$_\uparrow$	F1$_\uparrow$
Teacher [12]		0.548	0.0137	0.882	0.560
Audio-only$_{\text{Mono}}$		0.123	0.0607	0.608	0.160
Audio-only$_{\text{Stereo}}$		0.133	0.0571	0.623	0.161
[12]	MSE	0.124	0.0587	0.620	0.161
	Rank [6]	0.127	0.0582	0.624	0.163
	MTA [13]	0.128	0.0573	0.623	0.162
U-Net [14]	MSE	0.140	0.0573	0.649	0.166
	Rank [6]	0.137	0.0597	0.640	0.172
	MTA [13]	0.152	0.0577	0.650	0.173
	SAM$_{\text{Full}}$	**0.177**	**0.0535**	**0.671**	**0.180**
ViT [10]	MSE	0.142	0.0573	0.619	0.172
	Rank [6]	0.149	0.0564	0.629	0.171
	MTA [13]	0.142	0.0563	0.615	0.171
	SAM$_{\text{Full}}$	**0.175**	**0.0525**	**0.660**	**0.187**

Table 8: Comparison of 3D scene reconstruction accuracy on Replica with non-iid setting.

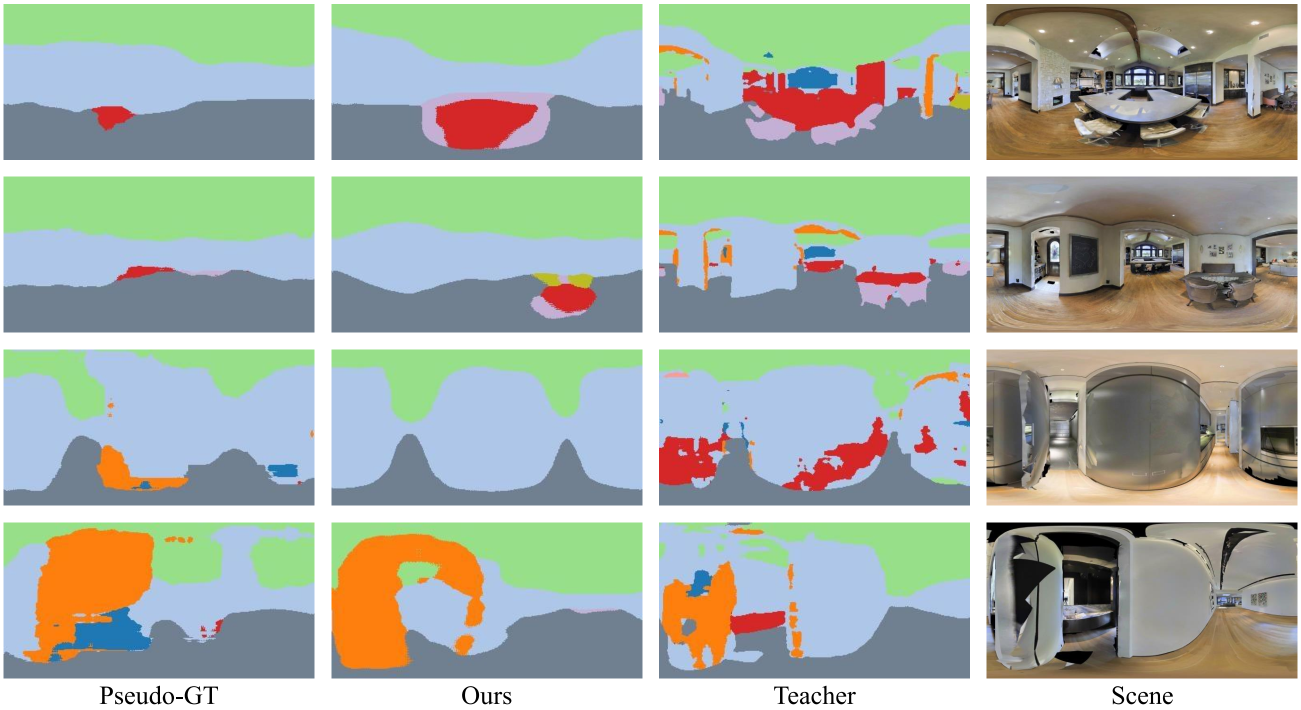

Figure 9: Qualitative examples of audio-based semantic segmentation on DAPS-Semantic test split.

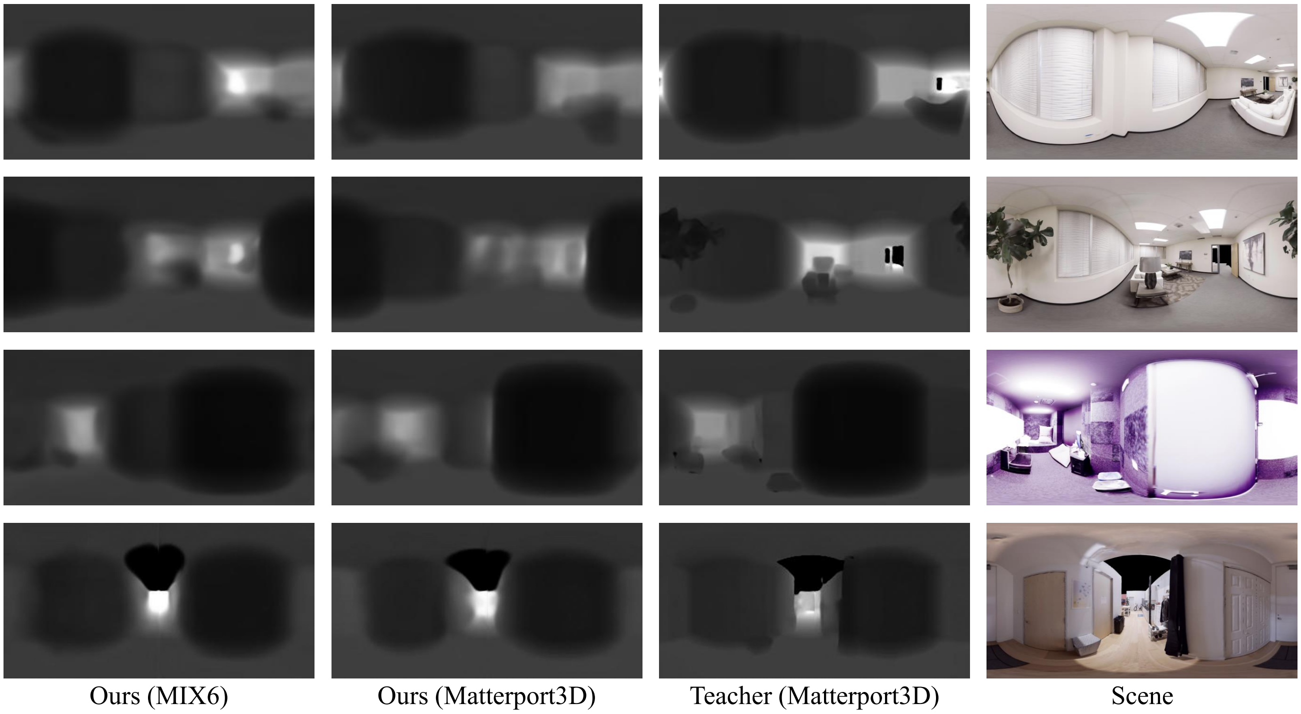

Figure 10: Qualitative examples of audio-based depth estimation on Replica.

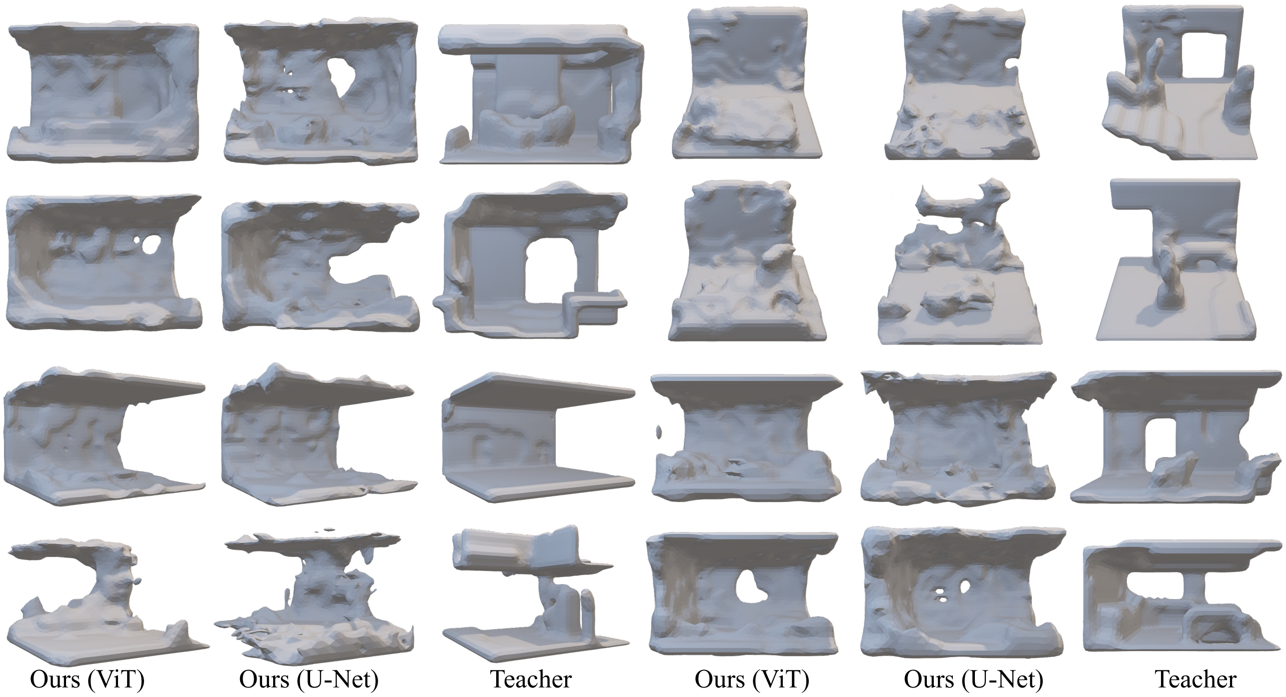

Figure 11: Qualitative examples of audio-based 3D scene reconstruction on DAPS-3D test split.

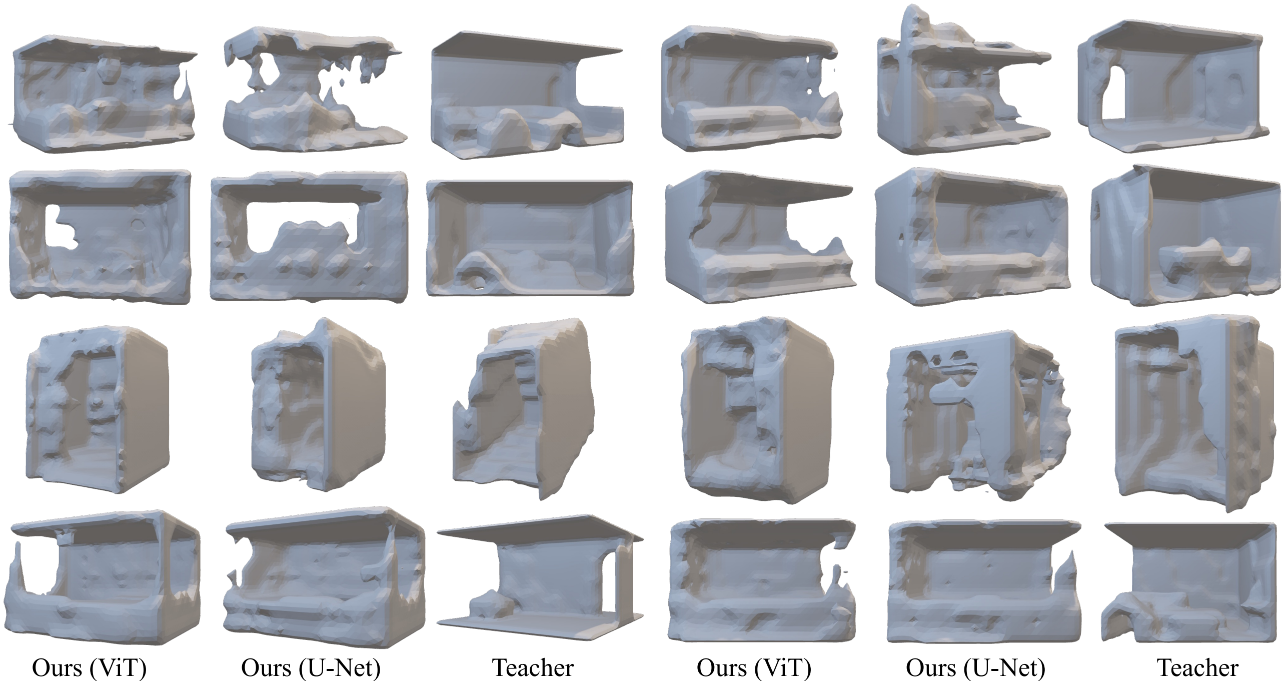

Figure 12: Qualitative examples of audio-based 3D scene reconstruction on Replica.

References

[1] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 1

[2] Hualie Jiang, Zhe Sheng, Siyu Zhu, Zilong Dong, and Rui Huang. Unifuse: Unidirectional fusion for 360 panorama depth estimation. *IEEE RA-L*, 2021. 1

[3] 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. In *3DV*, 2017. 1

[4] William E Lorensen and Harvey E Cline. Marching cubes: A high resolution 3d surface construction algorithm. *ACM SIGGRAPH*, 1987. 1

[5] Arun Balajee Vasudevan, Dengxin Dai, and Luc Van Gool. Semantic object prediction and spatial sound super-resolution with binaural sounds. In *ECCV*, 2020. 1

[6] Chuang Gan, Hang Zhao, Peihao Chen, David Cox, and Antonio Torralba. Self-supervised moving vehicle tracking with stereo sound. In *CVPR*, 2019. 1, 3

[7] Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, et al. The replica dataset: A digital replica of indoor spaces. *arXiv:1906.05797*, 2019. 2

[8] Changan Chen, Unnat Jain, Carl Schissler, Sebastia Vicenc Amengual Gari, Ziad Al-Halah, Vamsi Krishna Ithapu, Philip Robinson, and Kristen Grauman. Soundspaces: Audio-visual navigation in 3d environments. In *ECCV*, 2020. 2

[9] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE TPAMI*, 2020. 2

[10] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *ICCV*, 2021. 2, 3

[11] Jesper Haahr Christensen, Sascha Hornauer, and X Yu Stella. Batvision: Learning to see 3d spatial layout with two ears. In *ICRA*, 2020. 3

[12] Songyou Peng, Michael Niemeyer, Lars M. Mescheder, Marc Pollefeys, and Andreas Geiger. Convolutional occupancy networks. In *ECCV*, 2020. 3

[13] Francisco Rivera Valverde, Juana Valeria Hurtado, and Abhinav Valada. There is more than meets the eye: Self-supervised multi-object detection and tracking with sound by distilling multimodal knowledge. In *CVPR*, 2021. 3

[14] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015. 3