Acoustic NLOS Imaging with Cross-Modal Knowledge Distillation (Appendix)

Anonymous Author(s)
Affiliation
email@example.com

In this supplementary material, we describe a detailed implementation and experiments to determine the optimal conditions for cross-modal knowledge distillation (CMKD) in acoustic non-line-of-sight (NLOS) imaging. Additionally, we present the results of ablation experiments for the transfer learning of the translator and the data acquisition process. Finally, we include additional visualized results of the qualitative evaluation to supplement the information provided in the main paper.

A Implementation Details

We implemented the proposed acoustic NLOS imaging network using Pytorch and trained on the NVIDIA V100 GPU with 32GB of memory. The learning rate for the network was set to 0.0001, and the initial decision rates for the Adam optimizer ($\beta 1$ and $\beta 2$) were set to 0.5 and 0.999, respectively. The batch size was set to 32. The weights of the audio sub-network's translator were initialized with those of the pre-trained image sub-network, while the weights of the other networks were randomly initialized. Training for 100 epochs on the acquired dataset took approximately 2 hours.

The CMKD model is designed to predict depth maps of size 64x64, and thus the RGB images are also resized to 64×64. The raw waveform audio data is transformed into a 256×512 spectrogram using short-term Fourier transform with a window size of 512. The resulting multi-channel audio, which includes information about the x and y locations, is input to the model as 4D data of size 8×8×256×512.

B More Experimental Analysis

B.1 Optimal Condition for Knowledge Distillation

In this experiment, we investigate the optimal conditions for knowledge distillation in acoustic NLOS imaging.

We utilize a U-Net [Ronneberger *et al.*, 2015] translator to transfer knowledge, which is structured as an auto-encoder composed of an encoder and a decoder. In order to determine the most effective method, we evaluate three different scenarios of transferring knowledge from the encoder, the decoder, and the entire translator.

Additionally, we examine the use of various loss functions to learn the knowledge distillation process. Through a series of experiments, we determine the loss function that optimizes the performance of CMKD method.

(a)	KD	Rel(↓)	RMSE(↓)	δ_1(↑)	δ_2(↑)	δ_3(↑)
	Entire	3.240	**0.293**	52.0	61.6	68.4
	Decoder	3.964	**0.292**	50.8	61.4	68.4
	Encoder	**2.994**	0.293	**57.2**	**65.9**	**71.7**
(b)	Loss	Rel(↓)	RMSE(↓)	δ_1(↑)	δ_2(↑)	δ_3(↑)
	L1	3.707	**0.293**	46.8	57.4	64.8
	L2	3.265	0.299	45.6	56.1	63.7
	KL Div	**2.994**	**0.293**	**57.2**	**65.9**	**71.7**
(c)	Temp	Rel(↓)	RMSE(↓)	δ_1(↑)	δ_2(↑)	δ_3(↑)
	1	3.571	**0.287**	46.9	57.9	65.8
	4	3.909	0.318	47.1	57.2	64.2
	9	**2.994**	0.293	**57.2**	**65.9**	**71.7**

Table 1: Quantitative results to investigate the optimal conditions for knowledge distillation in acoustic NLOS imaging.

Furthermore, we explore the appropriate temperature value for the knowledge distillation process. The temperature value is an important hyper-parameter that controls the softness of the teacher model output probability distribution, which can affect the student model learning.

The results of experiments are summarized in Tab.1 and demonstrate that the best performance is achieved when only encoder knowledge is transferred using the KL div loss function and the temperature value of 9. Through these experiments, we present the optimal conditions for CMKD in acoustic NLOS imaging.

B.2 Transfer learning for audio translator

We present an audio translator that utilizes transfer learning by initializing the model with a pre-trained image translator. To evaluate the effectiveness of this approach, we also examine the performance of an audio translator initialized using random weights as a control.

The results of this experiment, presented in Tab. 2, demonstrate that transfer learning has a positive impact on the performance of the audio translator. A comparison of the results obtained from the two initialization methods illustrates the significant effect of transfer learning on CMKD method.

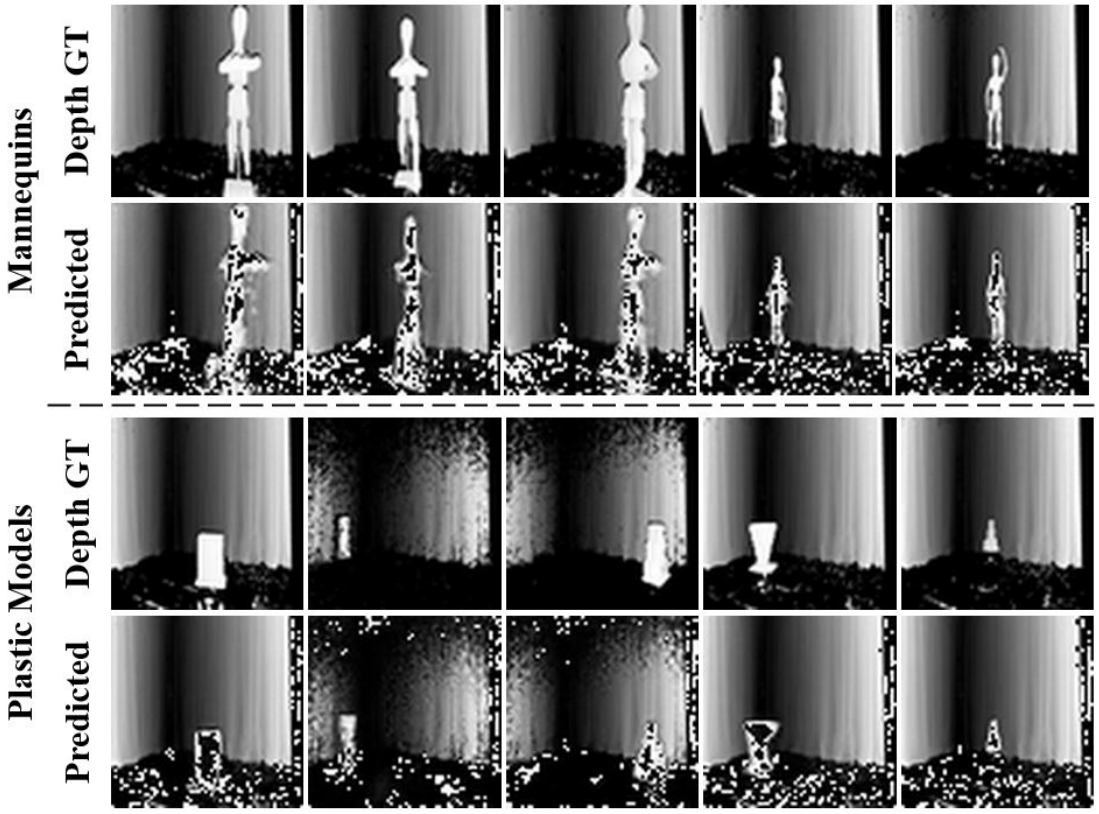

Figure 1: Additional visualized results of the reconstruction of trained objects using the CMKD model.

	Weight	Rel(↓)	RMSE(↓)	δ_1(↑)	δ_2(↑)	δ_3(↑)
(a)	Random	3.086	**0.285**	46.1	57.2	65.0
(b)	Pre-trained	**2.994**	0.293	**57.2**	**65.9**	**71.7**

Table 2: Results from the different initialization methods. (a) Performance of the audio translator initialized with random weights, (b) Performance of our proposed audio translator that and initialized with pre-trained image translator weights.

B.3 Data Acquired Process

Previous acoustic NLOS methods employ data acquisition techniques such as the sequential [Lindell *et al.*, 2019] or simultaneous [Jang *et al.*, 2022] emission method. The sequential emission method involves emitting linear chirp signals from eight speakers one at a time, while the simultaneous emission method emits signals concurrently from all speakers. While the simultaneous emission process is faster, the sequential method is considered to be more advantageous for analyzing the reflected signal as it reduces overlap between the emitted signals.

We acquired acoustic datasets with both sequential and simultaneous emission methods. By comparing the results obtained through these two methods, we aim to gain a deeper understanding of the trade-offs and advantages of each method and provide insights for acoustic NLOS imaging research. The data acquisition time per sample, including the movement of the speaker-microphone array, is 20 seconds for the simultaneous and 25 seconds for the sequential emission method, with a 20% difference. However, the reconstruction performance indicates that the sequential emission method is superior to the simultaneous emission method in all acoustic imaging models [Christensen *et al.*, 2020; Jang *et al.*, 2022], as shown in Tab. 3. Despite the longer collection time, we choose to use the sequential emission method due to its superior reconstruction performance.

Method	Approach	Rel(↓)	RMSE(↓)	δ_1(↑)	δ_2(↑)	δ_3(↑)
Simultaneous	Batvison	8.309	0.369	19.2	43.6	58.7
	HAE	11.540	0.465	44.0	53.2	58.6
	CMKD	10.884	0.442	44.5	54.2	59.5
Sequential	Batvison	5.311	**0.288**	44.3	56.5	64.2
	HAE	3.539	**0.288**	49.4	60.4	67.8
	CMKD	**2.994**	0.293	**57.2**	**65.9**	**71.7**

Table 3: Results from the different emission methods. The sequential emission method is superior in all acoustic imaging models.

C More Qualitative Results

We present additional visualized results of the reconstruction of trained and unseen objects using the CMKD model in Fig. 1 and 2.

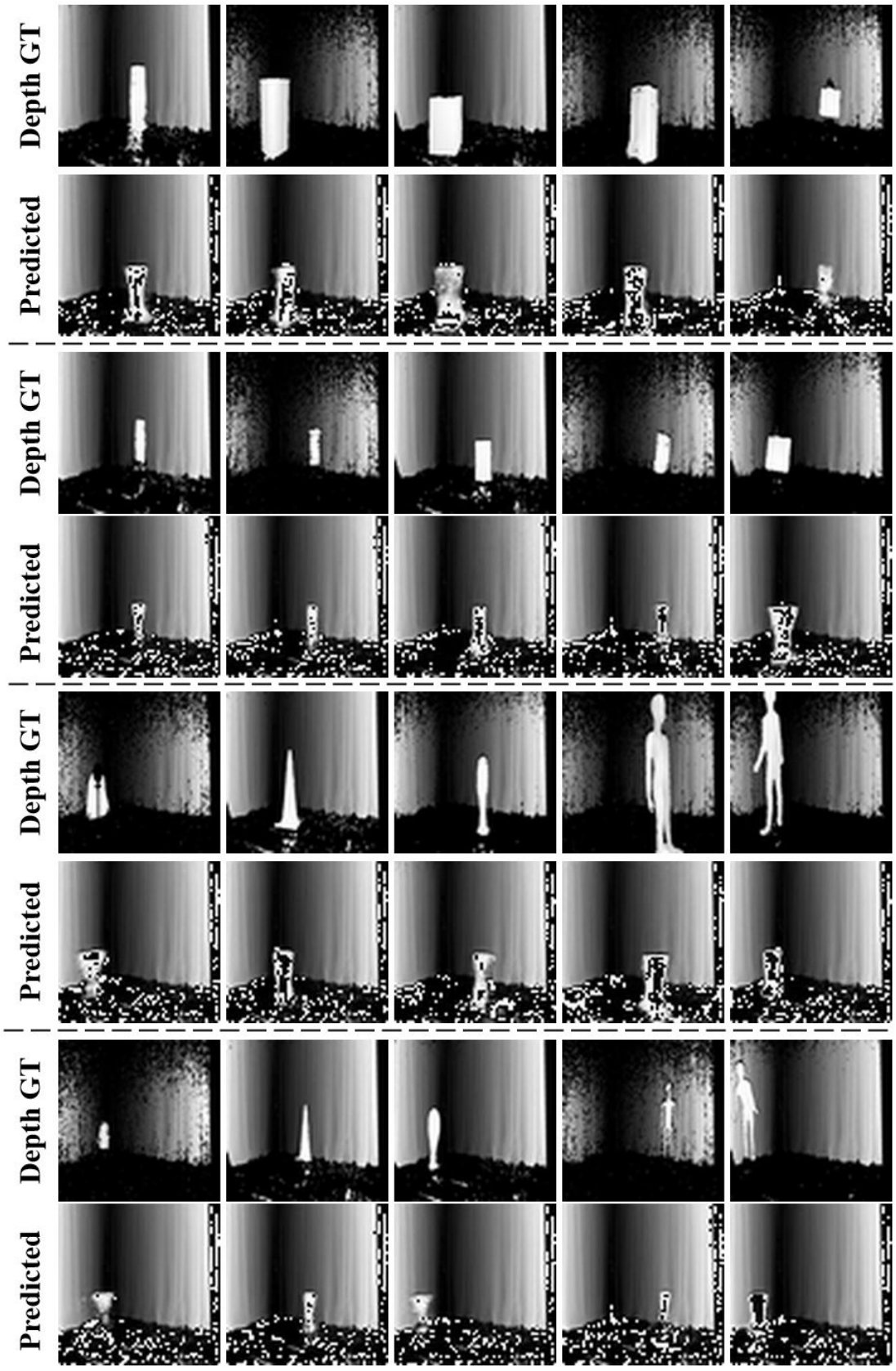

Figure 2: Additional visualized results of the reconstruction of unseen objects using the CMKD model.

References

[Christensen *et al.*, 2020] Jesper Haahr Christensen, Sascha Hornauer, and X Yu Stella. Batvision: Learning to see 3d spatial layout with two ears. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1581–1587. IEEE, 2020.

[Jang *et al.*, 2022] Seungwoo Jang, Ui-Hyeon Shin, and Kwangsu Kim. Deep non-line-of-sight imaging using echolocation. *Sensors*, 22(21):8477, 2022.

[Lindell *et al.*, 2019] David B Lindell, Gordon Wetzstein, and Vladlen Koltun. Acoustic non-line-of-sight imaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6780–6789, 2019.

[Ronneberger *et al.*, 2015] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.