Acoustic NLOS Imaging with Cross-Modal Knowledge Distillation

Ui-Hyeon Shin¹, Seungwoo Jang¹ and Kwangsu Kim²
¹Department of Artificial Intelligence, Sungkyunkwan University, Korea
²College of Computing and Informatics, Sungkyunkwan University, Korea
{shineh96, sewo, kim.kwangsu}@skku.edu

Abstract

Acoustic non-line-of-sight (NLOS) imaging aims to reconstruct hidden scenes by analyzing reflections of acoustic waves. Despite recent developments in the field, existing methods still have limitations such as sensitivity to noise in a physical model and difficulty in reconstructing unseen objects in a deep learning model. To address these limitations, we propose a novel cross-modal knowledge distillation (CMKD) approach for acoustic NLOS imaging. Our method transfers knowledge from a well-trained image network to an audio network, effectively combining the strengths of both modalities. As a result, it is robust to noise and superior in reconstructing unseen objects. Additionally, we evaluate real-world datasets and demonstrate that the proposed method outperforms state-of-the-art methods in acoustic NLOS imaging. Our code, model, and data are available at https://github.com/shineh96/Acoustic-NLOS-CMKD.

1 Introduction

Non-line-of-sight (NLOS) imaging [Kirmani et al., 2009] is a method for reconstructing objects or scenes that are hidden from the line-of-sight of an observer. Conventional NLOS imaging methods [Velten et al., 2012; Heide et al., 2014; O’Toole et al., 2018] primarily utilize optical systems in order to infer the properties of hidden scenes. These are achieved by analyzing indirect measurements, such as reflections of optic waves. However, acoustic signals can also be used for NLOS imaging, providing an alternative approach to the analysis of optical signals. Acoustic signals are immune to interference or noise from external sources, such as light or radio frequency radiation. Furthermore, the audible frequency signal exhibits robustness to noise within a specific frequency band, owing to its wide frequency range of 20 Hz to 20 kHz. This makes acoustic NLOS systems more robust and reliable in noise environments, or in situations where the reflections of the optical waves may be distorted or attenuated. In contrast, optical NLOS systems may be affected by noise, which can reduce the quality of the reconstructed image.

Recently, NLOS imaging methods that utilize acoustic characteristics have been proposed. [Lindell et al., 2019a] proposed a physical model for analyzing acoustic time-of-flight, inspired by seismic imaging. However, NLOS systems typically measure three-bounce reflected signals, as shown in Fig. 1. These signals have low signal intensity, a long travel distance, and high levels of environmental noise. Furthermore, the measurements may be affected by ambient noise, interference, or multipath effects, which can degrade the accuracy and reliability of the time-of-flight estimates. As a result, this approach has only been verified with data collected in a space that is isolated with acoustic foam panels and does not reflect acoustic signals other than those of the relay wall.

To address the limitations of the physical model, [Jang et al., 2022] proposed an end-to-end deep learning model that reconstructs the depth map by extracting the features of hidden scenes from the relative intensity and the arrival time delay of the reflected signal. The model utilizes an encoder with a hierarchical structure to extract acoustic signals from multi-
channel audio and the reconstruct hidden scenes in a space where no soundproofing system has been implemented. However, the model is limited in its ability to reconstruct unseen objects that are out of distribution with respect to the trained objects.

In general, knowledge distillation [Hinton et al., 2015] has been shown to improve the generalization performance of a target student model by transferring the knowledge of a verified teacher model [Stanton et al., 2021]. Additionally, several studies [Aytar et al., 2016; Albanie et al., 2018; Gan et al., 2019; Valverde et al., 2021] have demonstrated that knowledge distillation between different modalities, such as from image to audio, can further enhance the performance of the target model. Based on these findings, we design a model that is optimized for acoustic NLOS imaging that is intended to be robust to noise and capable of reconstructing unseen objects. To achieve this, we propose a cross-modal knowledge distillation (CMKD) approach that transfers the knowledge of a well-trained image network to an audio network.

The utilization of CMKD allows the strengths of each modality to be used optimally. Image data faithfully represent visual details and spatial information, whereas audio effectively capture dynamic information and potentially useful temporal information. By combining these strengths, the model could achieve better performance than by using either modality alone [Zhao et al., 2018; Gao et al., 2020]. Furthermore, this method enables the model to better generalize to unseen objects and makes the target network robust to noise [Sarfraz et al., 2021].

The CMKD framework consists of an image teacher network and an audio student network shown in Fig. 2. The image teacher network is initially trained to perform the transformation of an RGB image into a depth map. Subsequently, the audio student network is trained to convert multi-channel audio to a depth map, and to leverage the distilled knowledge from the frozen image teacher network. During inference, the audio student network is able to reconstruct the depth map of a hidden scene using only reflected acoustic signals as input, without any additional image information.

To facilitate this task, we collect a large dataset of 3,600 corresponding frames that consist of RGB images, depth maps, and multi-channel audio. We also construct an acoustic system with eight speaker and microphone arrays and collect 64 channels of reflected signals by transmitting and receiving audible signals (20 Hz to 20 kHz) in a space where no soundproofing system has been implemented. We use this self-collected experimental data to confirm the robustness of our model to noise generated in real-world scenarios.

We compare the performance of our approach with state-of-the-art methods using acquired data. We demonstrate superior performance in reconstructing both trained and unseen objects. We also present detailed ablation studies to highlight the significance of the proposed techniques. The main contributions of this work are as follows:

- To the best of our knowledge, this is the first instance where CMKD has been applied to NLOS imaging in general, not just in the acoustic domain.
- We collect a new acoustic NLOS dataset and make it available to the public. We hope that this dataset will contribute to the advancement of research in the field of acoustic NLOS.
- Our model demonstrates robustness to real-world noise and enhances the generalization performance on unseen objects, and it outperforms current state-of-the-art models.

2 Related Work
2.1 NLOS Imaging

NLOS imaging has numerous potential applications, including autonomous driving, medical imaging, and rescue operations [Maeda et al., 2019]. A variety of hardware systems, such as pulse lasers and high-resolution detectors [Velten et al., 2012; Liu et al., 2020; Wu et al., 2021], time-of-flight cameras [Heide et al., 2014; Kadamb et al., 2016], conventional cameras [Chen et al., 2019; Henley et al., 2020], LiDAR systems [Zhu and Cai, 2022], and speaker-microphone arrays [Lindell et al., 2019a; Jang et al., 2022], have been used for NLOS imaging. Additionally, several methods have been proposed, including time-of-flight-based models [Velten et al., 2012; Heide et al., 2014] that use directivity and wave-based models [Lindell et al., 2019b] that use diffraction. However, NLOS imaging is an ill-posed problem with a low signal-to-noise ratio, due to the fact that it relies on the analysis of three-bounce reflected signals [Geng et al., 2021]. This can make it challenging to achieve high-quality reconstruction of the hidden scene.

To address this problem, several NLOS imaging methods that use deep learning [Chen et al., 2019; Grau Chopite et al., 2020; Shen et al., 2021] have been proposed. These methods have been successful in reconstructing hidden scenes by distinguishing noise and extracting meaningful features. However, it is important to note that the performance of deep
learning models heavily relies on the quantity and quality of the dataset. In particular, it is difficult to construct large datasets using optical equipment for NLOS imaging. This is primarily due to the directivity of light, which results in a long collection time of 1 - 5 minutes per sample using point-by-point scanning. As a result, most deep learning methods rely on synthesized data for training and evaluation, as it can be impractical to collect large amounts of real-world data. In contrast, we reduce the collection time to 25 seconds by using an acoustic system that can scan hidden spaces at once. This enables the collection of a larger and higher-quality dataset, which is crucial for the performance of deep learning models.

3 Methodology

In this section, we provide a detailed description of the overall framework and the role of each component, as well as the knowledge transfer method between the two modalities and the loss function used for network learning.

The goal of this framework is to reconstruct the depth map for the hidden scene by transferring knowledge from an image modality to an audio modality. To achieve this, we use an RGB image as the teacher modality and multi-channel audio as the student modality. We employ a two-phase approach, where the first phase involves training the teacher network to transform an RGB image to a depth map. Then, in the second phase, the weights of the well-trained teacher network are frozen, and a student network is trained to convert multi-channel audio to a depth map using the distilled knowledge supplied by the teacher network.

3.1 Cross-Modal Knowledge Distillation

In our approach, we use this method to transfer the knowledge of a well-trained RGB image to depth network to the audio to depth network, with the goal of improving the reconstruction performance of the audio network. During training, we learn both the image and audio modalities, but during inference, we only use the audio modality. This approach allows us to effectively transfer the knowledge of the image network to the audio network, resulting in improved performance.

To facilitate CMKD, the translators of the two sub-networks are designed to have the same structure. We compare three cases of transferring knowledge from the image teacher network to the audio student network: encoder knowledge, decoder knowledge, and whole network knowledge. The results show that transferring only the knowledge of the encoders leads to the greatest improvement in the performance of the audio network. The detailed results of this experiment can be found in the supplementary material. Based on these findings, we present optimal conditions for CMKD in acoustic NLOS imaging.

3.2 Network Architecture

The network architecture consists of two main components: an image teacher network and an audio student network.

Image Teacher Network

The image teacher network is a translator that converts RGB images into depth maps. We adopt a U-Net [Ronneberger et al., 2015] structure auto-encoder as the translator network. The U-Net has been shown to perform well on the task of...
monocular depth estimation [Alhashim and Wonka, 2018], which involves converting each pixel of an RGB image to a depth value.

The U-Net translator consists of an encoder that extracts features from an RGB image and a decoder that reconstructs the latent vector as a depth map. The encoder and decoder are symmetrical, and the high-dimensional information from the encoder is transmitted to the decoder through skip connections. This image network learns the knowledge that is required to convert RGB images to depth maps.

Audio Student Network
The audio student network consists of three main components: a feature extractor optimized for multi-channel audio input, a translator that converts the extracted audio features to a depth map, and a discriminator that distinguishes whether the estimated depth map is real or fake. The feature extractor is responsible for extracting meaningful features from the multi-channel audio input, these features are then passed to the translator. The translator uses these features to reconstruct the depth map of the hidden space. The discriminator is used to evaluate the quality of the reconstructed depth map and distinguish between real and fake examples.

The audio network feature extractor is designed specifically to manage multi-channel audio data that are acquired from various locations. The audio data are acquired using an 8×8 grid of vertically arranged speaker-microphone pairs that move horizontally. To extract features from the 1D time series data, we apply a short-time Fourier transform to convert the data into a 2D spectrogram having dimensions of 256×512. The resulting 4D audio data (8×8×256×512) are input into the network and passed through eight encoding blocks that extract features using 3D convolution operations [Tran et al., 2015] and two fully connected layers that transform the latent vector to the input form for the next network. Each encoding block consists of a 3D convolutional layer, a 3D batch normalization layer, and a ReLU activation function. This network effectively extracts features from the 4D audio data while preserving the location information.

The translator in the audio network has the same structure as the image network, which allows for the transfer of knowledge from the image network to the audio network. This structure, which is based on the RGB image to depth map translator, helps to improve the reconstruction performance of the audio network. In addition, the student translator is initialized with the pre-trained weights of the teacher network in order to accelerate learning and further improve reconstruction performance.

We adopt the discriminator structure from Pix2Pix [Isola et al., 2017]. The discriminator serves the purpose of distinguishing whether the estimated depth map is real or fake. The discriminator aligns the distribution of the prediction depth map with the ground truth depth map.

3.3 Objective
Image Teacher Network
The image network is trained using only the depth Loss, which is the pointwise L1 error between the estimated depth map and the actual depth map. The objective of the image network is as follows:

\[ G_t^* = \min_{G_t} \mathcal{L}_{Depth}(G_t), \]  

where, \(G_t\) is a teacher network generator that translates the RGB image to the depth map.

Audio Student Network
The audio network employs knowledge distillation to enhance the performance of the conditional adversarial network for audio to depth map translation. Therefore, it is trained by integrating the loss for the conditional adversarial network with the loss for the knowledge distillation. We utilize a conditional adversarial network loss based on the Batvion [Christensen et al., 2020] and we measure the depth map reconstruction error using the pointwise L1 error. The GAN loss is determined by the least-squares loss [Mao et al., 2017]. In order to align the audio network with the image network, the distance between the feature map distributions of each translator encoding block should be minimized. Our network is designed to minimize this distance as measured by the Kullback-Leibler divergence (KL divergence) [Hinton et al., 2015]. The objective of the audio network is as follows:

\[ G_s^* = \min_{G_s} \max_{D_s} \frac{1}{2} \mathcal{L}_{GAN}(D_s) + \mathcal{L}_{GAN}(G_s) + \alpha \mathcal{L}_{Depth}(G_s) + \beta \mathcal{L}_{KL}(G_s), \]  

where, \(G_s\) is the generator of the student network, and \(D_s\) is the discriminator of the student network. \(\alpha\) and \(\beta\) are balancing weights. We set \(\alpha\) to 100 and \(\beta\) to 0.01.

4 Experiment
In this section, we describe the data acquisition system for acoustic NLOS imaging and the details of the experimental setup using the acquired dataset. We then evaluate the performance of CMKD approach for NLOS imaging and compare it with state-of-the-art methods for both LOS and NLOS acoustic imaging. We demonstrate the superiority for unseen object reconstruction and present detailed ablation studies to highlight the contributions of techniques in our method.

4.1 Data
The data used in the experiments and evaluations were self-collected and are representative of real-world scenarios. Using self-acquired data, rather than synthetic or simulated data, enhances the external validity of the results and makes it more likely that the results can be generalized to real-world scenarios. In this subsection, we describe the experimental setup, data acquisition equipment, and processes used in this study on acoustic NLOS imaging.

We conduct the experiments in a space without sound-proofing. The experimental setup includes an occluder that separates the scanning space from the hidden space shown on the left side of Fig. 4. The right side of Fig. 4 illustrates the configuration of the acoustic system. The system consists of eight sets of speakers and microphones, an audio interface, and a power amplifier. A translation stage is positioned at a 45-degree angle to the relay wall, to move the speaker-microphone array horizontally.
In the acoustic data acquisition process, we employ a sequential emission method. This method emits linear chirp signals by eight speakers in the audible frequency range (20 Hz to 20 kHz), each lasting for 0.1 seconds. To acquire the acoustical data, eight microphones were placed at intervals of 10 cm, with the speakers emitting linear chirp signals sequentially for a total of 0.8 seconds. The reflected signal is recorded simultaneously on all eight microphones for a duration of 0.9 seconds at a sampling rate of 48 kHz, where the time required for the last emitted signal to be reflected back is 0.1 seconds. The acoustic data were then collected at eight points, with the speaker-microphone array moving horizontally at intervals of 5 cm. Along with the acoustic data, we also acquired RGB images and depth maps as the ground truth for the hidden scene.

We acquired data using 30 different kinds of objects, including mannequins, plastic models, and other objects. The mannequins were posed differently for each class, and the plastic models were made to have various shapes such as hexahedrons and pyramids. Other objects included items such as paper boxes, backpacks, and plastic signs. Fig. 5 shows some examples of the target objects that were used for data acquisition. Each object is acquired 120 times at different angles and positions, resulting in a total of 3,600 time-synchronized RGB images, as well as depth maps and multi-channel audio.

4.2 Experimental Settings

Data Split

During the training process, we utilize only the mannequin and plastic model data. The data for the training objects are divided into 1920 samples for training, 240 samples for validation, and 240 samples for testing. The remaining objects, which are not used for training, are utilized to evaluate the model performance on unseen object reconstruction with a total of 1200 data samples.

Evaluation Metric

To evaluate the performance of methods for the depth map reconstruction of hidden scenes, we utilize metrics commonly used in depth estimation tasks [Alhashim and Wonka, 2018]. It is important to note that all data were acquired with the same background, and the size of the object region is only about 10% of the background region on average. Therefore, if the entire depth map is evaluated, a network that performs well on estimating the depth of the background may appear superior to a network that accurately predicts the depth of the target object. To address this issue, we evaluate the depth map reconstruction error for the object region only, excluding the background.

4.3 Baselines

We compare the performance of CMKD method with both LOS and NLOS acoustic imaging approaches. A physical model [Lindell et al., 2019a] reconstructs a hidden scene based on the analysis of acoustic time-of-flight. A Batvision [Christensen et al., 2020] is a state-of-the-art deep learning method for LOS acoustic imaging, which consists of an audio feature extractor, an auto-encoder, and a discriminator. A hierarchical audio encoder (HAE) [Jang et al., 2022] is a deep learning method for NLOS acoustic imaging that extracts audio features through the HAE that considers the location characteristics of multi-channel audio.

4.4 Experimental Results

We conduct experiments on both trained and unseen objects from the acquired dataset. We compare the performance of our method with several state-of-the-art acoustic imaging baseline methods using both quantitative and qualitative evaluation.

Quantitative Evaluation

In order to perform a quantitative evaluation, we evaluate the reconstruction error for only the object region to use depth estimation metrics. The physical model has limited capability for high resolution depth map reconstruction, which makes it difficult to directly compare it with other models. Therefore, we compare quantitative evaluations of proposed model with those of other baseline models.

In Tab. 1, CMKD shows the best performance in terms of quantitative evaluation on both trained and unseen objects. In particular, the threshold accuracy ($\delta_i$), which represents accuracy within certain tolerances, of our method shows a 10 - 20% improvement over that of other methods. Although the RMSE of our model is slightly higher than that of other models, the difference is small, ranging from 1 - 5%. Other methods tend to blur areas where objects are expected to be, as
Table 1: Results of the quantitative evaluation. The left side represents the results for trained objects, and the right side represents the results for unseen objects. Rel is the relative error, and RMSE is the root mean square error. \( \delta_i \) is the percentage of pixels for which the depth estimates are within a certain range of the true depths. "↑" means that higher is better and "↓" means that lower is better.

<table>
<thead>
<tr>
<th>Approach</th>
<th>Trained Objects</th>
<th>Unseen Objects</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Rel(↓) RMSE(↓)</td>
<td>( \delta_1 )↑ ( \delta_2 )↑ ( \delta_3 )↑</td>
</tr>
<tr>
<td>Batvision</td>
<td>5.311 0.288</td>
<td>44.3 56.5 64.2</td>
</tr>
<tr>
<td>HAE</td>
<td>3.539 0.288</td>
<td>49.4 60.4 67.8</td>
</tr>
<tr>
<td>CMKD (Ours)</td>
<td>2.994 0.293</td>
<td>57.2 65.9 71.7</td>
</tr>
</tbody>
</table>

Figure 6: Visualized results of the qualitative evaluation. The left side shows the results for trained objects, and the right side shows the results for unseen objects. CMKD model is able to clearly reconstruct the shape of both trained and unseen objects. In contrast, the baseline models either produce blurry reconstructions or fail to detect the unseen objects.

In this subsection, we qualitatively evaluate the performance of CMKD framework for acoustic NLOS imaging. Fig. 6 shows the visualized results for depth map reconstruction for trained and unseen objects, respectively.

Our experiments are conducted in a non-soundproofed environment with ambient noise and overlapping reflections, which can be challenging for the physical model. However, deep learning models, including our model, accurately reconstruct the background due to their ability to learn from data with the same background.

In the case of trained objects, both Batvision and HAE approximate the location of hidden objects and reconstruct their shapes. However, these baseline models sometimes fail to accurately detect object locations and the shapes of their reconstructions are not always clear. In contrast, CMKD model accurately estimates both the shape and distance of the hidden object, and it accurately detects the area where the object is located.

Additionally, we evaluate the generalization performance of these models through experiments on unseen objects. While most deep learning-based methods detect the areas where hidden objects are located, Batvision struggles to accurately estimate object shapes and tends to reconstruct blurry depth maps. HAE reconstructs box-shaped objects relatively well, but performs poorly on untrained objects of other shapes. In contrast, CMKD model accurately reconstructs both the position and shape of the object thanks to the transfer of knowledge from the image teacher network, which is not utilized by the other methods.

Other deep learning baselines rely on the pixel-wise loss. However, in some cases, using the pixel-wise loss function may lead to a blurry reconstruction because the model is unable to capture fine-grained details or sharp edges in the image. This can occur if the model does not have enough capacity or if the training data are not representative of the test data. In contrast, our model utilizes knowledge distillation...
Ablation Study

<table>
<thead>
<tr>
<th>Extractor</th>
<th>KD</th>
<th>Rel(%)</th>
<th>RMSE(%)</th>
<th>δ₁(%)</th>
<th>δ₂(%)</th>
<th>δ₃(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a) 3D CNN</td>
<td>X</td>
<td>7.888</td>
<td>0.399</td>
<td>31.6</td>
<td>42.0</td>
<td>49.1</td>
</tr>
<tr>
<td>(b) 2D CNN</td>
<td>O</td>
<td>8.027</td>
<td>0.397</td>
<td>36.8</td>
<td>46.9</td>
<td>53.7</td>
</tr>
<tr>
<td>(c) HAE</td>
<td></td>
<td>7.479</td>
<td>0.396</td>
<td>36.8</td>
<td>46.8</td>
<td>53.4</td>
</tr>
<tr>
<td>(d) 3D CNN</td>
<td></td>
<td>7.094</td>
<td>0.392</td>
<td>40.0</td>
<td>49.9</td>
<td>56.2</td>
</tr>
</tbody>
</table>

Table 2: Results of ablation studies. (a) Performance when knowledge distillation is not applied to the audio network structure. (b), (c) Performances when the audio feature extractor is replaced with a 2D CNN and a hierarchical 2D CNN, respectively. (d) Our method using a 3D CNN feature extractor and incorporating knowledge distilled from the image network significantly improves the reconstruction of hidden objects in acoustic NLOS imaging. These findings confirm the effectiveness of the techniques and structures implemented in the proposed model.

5 Conclusion

In this paper, we propose a method for improving the performance of acoustic NLOS imaging systems. While previous approaches to acoustic NLOS imaging have encountered limitations, such as vulnerability to noise and difficulty in reconstructing unseen objects, our method uses CMKD to transfer knowledge from a well-trained image network to an audio network. This enables the resulting model to be robust to noise and to enhance the generalization performance on unseen objects. Our experimental results show that CMKD method outperforms state-of-the-art methods in acoustic NLOS imaging and demonstrates superior performance in reconstructing unseen objects. Additionally, the results of the ablation studies demonstrate the suitability of the techniques and structures implemented in the proposed model for acoustic NLOS imaging. Overall, we provide a promising solution for acoustic NLOS imaging, and has potential for various practical applications in the future.

Acknowledgments

This work was partly supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.2021-0-02068, Artificial Intelligence Innovation Hub) and Korea Internet & Security Agency (KISA) grant funded by the Korea government(MSIT) (No. RS-2023-00231200, Development of personal video information privacy protection technology capable of AI learning in an autonomous driving environment).

References


[Christensen et al., 2020] Jesper Haahr Christensen, Sascha Hornauer, and X Yu Stella. Batvision: Learning to see


