This ICCV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

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

Heeseung Yun,* Joonil Na,* Gunhee Kim
Seoul National University
{heeseung.yun, joonil}@vision.snu.ac.kr, gunhee@snu.ac.kr
https://github.com/hs-yn/DAPS

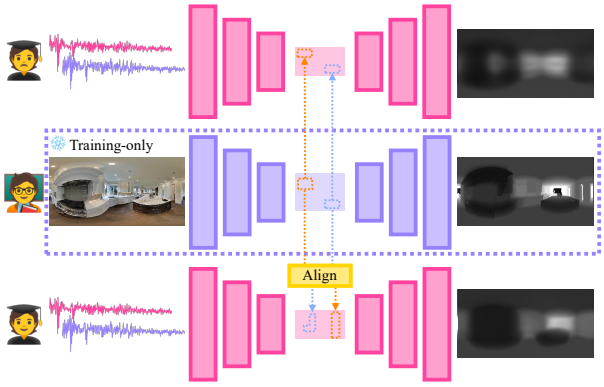
Abstract

Sound can convey significant information for spatial reasoning in our daily lives. To endow deep networks with such ability, we address the challenge of dense indoor prediction with sound in both 2D and 3D via cross-modal knowledge distillation. In this work, we propose a Spatial Alignment via Matching (SAM) distillation framework that elicits local correspondence between the two modalities in vision-to-audio knowledge transfer. SAM integrates audio features with visually coherent learnable spatial embeddings to resolve inconsistencies in multiple layers of a student model. Our approach does not rely on a specific input representation, allowing for flexibility in the input shapes or dimensions without performance degradation. With a newly curated benchmark named Dense Auditory Prediction of Surroundings (DAPS), we are the first to tackle dense indoor prediction of omnidirectional surroundings in both 2D and 3D with audio observations. Specifically, for audio-based depth estimation, semantic segmentation, and challenging 3D scene reconstruction, the proposed distillation framework consistently achieves state-of-the-art performance across various metrics and backbone architectures.

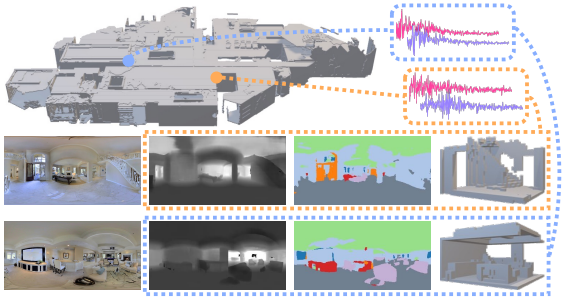
1. Introduction

Humans can get a good grasp of various information about surroundings with hearing without seeing, like the size of a room or the location of an active alarm. A long line of research has analyzed such intriguing abilities of humans based on interaural differences [1, 2] or brain activation with respect to spatially aligned audio-visual inputs [3, 4], to list a few. Accordingly, there is an emerging interest in teaching neural network models for spatial reasoning without seeing. Such models that spatially perceive the surroundings from sound can be utilized in various environments that are critical for privacy preservation or visually ill-posed (*e.g.*, low illumination or occlusion) [5, 6, 7, 8].

*Equal Contribution



(a) Mitigating inconsistencies in Cross-Modal Distillation



(b) Dense Auditory Prediction of Surroundings

Figure 1: key idea of our approach. (a) For vision-to-audio cross-modal distillation, instead of direct distillation between geometrically inconsistent modalities, we spatially align the latent feature maps of students with those of teachers. (b) Using auditory input only, we perform three dense predictions of surroundings: depth estimation, semantic segmentation, and 3D scene reconstruction.

Since predicting visual properties directly from audio is challenging, cross-modal knowledge distillation [9] is often utilized, *i.e.*, teaching audio models with the guidance of visual models. Visual models can make precise predictions about the image of the surroundings, like the location of objects or the depth of a scene. Thus, using visual models as the teacher, audio models can learn how to predict visual properties in a scene from sound inputs. This cross-modal knowledge distillation has been successfully applied

to make audio models predict *sparse* attributes, *e.g*., vehicle tracking [5] or indoor navigation [7]. However, it remains challenging to make *dense* visual predictions about the surroundings from audio.

One of the core challenges behind the dense prediction with audio is to identify fine-grained attributions of the output. In other words, humans can intuitively make sense of the room layout by hearing, but have difficulty in explaining which bandwidths or timeframes are responsible for their perception. Unlike distilling an RGB image teacher for a thermal image student that is geometrically consistent up to the pixel level, there is no obvious one-to-one alignment between image and audio. Hence, it is not feasible to determine which part of the audio spectrogram corresponds to which region of the surrounding. While using multiple intermediate features of a teacher model as a guide can still be beneficial [5, 8], it may not be possible to solve the underlying local correspondence problem between the two heterogeneous modalities.

In this work, we are the first to address the dense indoor prediction of omnidirectional surroundings in both 2D and 3D with audio observations. To resolve the inconsistency problem, we propose a novel Spatial Alignment via Matching (SAM) distillation framework. SAM matches local correspondences between the two heterogeneous features by making use of learnable spatial embeddings in several layers of the audio student model, combined with loose triplet-based learning objectives. We retain a set of learnable spatial embeddings to capture spatially varying information of each layer, which are pooled and integrated with initial audio features for alignment. This allows us to resolve inconsistencies even when the shape of the audio input does not match that of the desired output, making it trivially extendable to a challenging scenario like audio-to-3D distillation.

To comprehensively evaluate the performance of our method, we curate a new benchmark for audio-based dense prediction of surroundings based on Matterport3D [10] and SoundSpaces [7]. We collect 15.8K indoor scene multimodal observations with task-specific annotations for audio-based depth estimation, semantic segmentation, and 3D scene reconstruction. In dense auditory prediction tasks spanning from 2D to 3D, our framework consistently improves the performance by a wide margin, which is validated on multiple architectures like U-Net [11], DPT [12], and ConvONet [13]. Also, qualitative results demonstrate that our approach can precisely predict the structure of the indoor environment with hearing without seeing.

2. Related Works

Indoor Multimodal Scene Analysis. Extensive research has been conducted to understand indoor surroundings for given various inputs. Using monocular images as input, many visual scene understanding tasks like depth estimation, semantic segmentation, and surface normal estimation have been studied [14, 10, 15]. In addition, 3D-based methods for semantic segmentation, object recognition, and floorplan reconstruction have been proposed with voxel or mesh-based representations [10, 15, 16, 17]. When performing such tasks, combining different modalities as inputs is proven to be effective, such as RGB with depth information for semantic segmentation [18] or voxels with point clouds for 3D segmentation [19]. Recently, 2D vision-language models are successfully employed for open-vocabulary 3D scene understanding [20, 21].

There has been a surge of interest in combining audio and visual signals to tackle visual or acoustic tasks in indoor environments. Some prior works generate binaural audio [22] or scene-aware auditory responses [23, 24] by utilizing visual surroundings as a reference. Binaural audio is simulated from a 3D scene for audio-visual embodied navigation [7, 25]. Audio signals can help improve performances in visual tasks like floorplan reconstruction [26] and depth estimation of normal field-of-views [27, 28].

Cross-modal Knowledge Distillation. Knowledge distillation [29] aims at transferring knowledge from a teacher model to a student model by minimizing the distances between the two logit distributions. Cross-modal distillation [9] enhances this transfer by ensuring that the intermediate features of the student model align with those of the teacher model when their input modalities are different. Distillation between different modalities can improve the robustness of prediction under diverse conditions, such as utilizing depth sensors in student models by distilling object detection, action recognition, or semantic segmentation models [30, 31, 32]. Likewise, Vobecky *et al*. [33] leverage LiDAR and image inputs to generate spatially consistent object proposals for semantic segmentation.

Cross-modal distillation can be applied to the scenarios where no explicit correspondence exists between the two modalities. Zhao *et al*. [34] use a student model with radio signals for human pose estimation via distillation. Roheda *et al*. [35] conditionally utilize noisy observations of available sensors like seismic sensors to enhance image quality. Also, audio-only and image-only teachers can teach a video-only student model via shared latent embedding [36] or long short-term memory networks [37] for better classification. Other examples include knowledge transfer of speech models for visual lip reading [38, 39] or visual captioning models for audio captioning [40].

Spatial Reasoning with Sound. Sound contains valuable information for spatial reasoning. Embodied agents can navigate indoor environments by relying solely on auditory input [7], and their exploration behavior can be promoted by referring to auditory feedback [41]. Other prior works focus on the spatial localization of audio sources [42], 3D face synthesis from speech [43], and depth

estimation on a robot [44, 45, 46]. Sound-only models can benefit from the cross-modal distillation of visual teacher models for fine-grained spatial understanding. Vision-to-audio knowledge distillation has shown compelling performance in vehicle localization [5, 8], obstacle detection [47], and collision probability estimation [48]. However, prior works are limited to the sparse prediction of the surrounding environment (*e.g.*, bounding boxes), while the dense prediction remains challenging.

Closest to our approach is Binaural SoundNet [6, 49], as it improves outdoor dense prediction performance through the cross-modal distillation of multiple tasks. However, our work has three significant differences. First, we perform indoor semantic segmentation and 3D scene reconstruction from audio as new dense prediction tasks. Second, SoundNet does not consider feature-level alignment, while our method hierarchically leverages spatial alignment via matching for fine-grained vision-to-audio distillation. Finally, instead of designing a new architecture for modeling audio inputs [5, 49] or forcing specific input representations [6, 8], we take the audio input as is and adapt off-the-shelf vision models for audio-based dense prediction.

3. Approach

Our goal is to predict various dense properties of indoor surroundings without visual input by leveraging binaural audios, *e.g.*, depth, semantic labels, and 3D structures. To this end, we present a framework for vision-to-audio knowledge distillation that does not rely on specific architecture and entails the alignment of heterogeneous features, as shown in Fig. 2. Given a pre-trained visual teacher, we aim to train an audio student model using paired audio-visual observations as training data.

We start by reviewing the basics of vision-to-audio knowledge distillation and the challenges in adapting such methods for dense auditory prediction of surroundings (§3.1). Next, we explain the proposed spatial alignment via matching distillation (§3.2). Finally, we outline training and inference procedures shared among different tasks (§3.3). Commonly used variables are defined as follows.

$a_{\text{in}}, v_{\text{in}}$	Audio, visual input ($\mathbb{R}^{W' \times H' \times 2}, \mathbb{R}^{W \times H \times 3}$)
$a_{\text{out}}, v_{\text{out}}$	Audio, visual prediction output ($\mathbb{R}^{W \times H}$)
a_i, v_i	Features at layer i ($\mathbb{R}^{A_i \times C}, \mathbb{R}^{V_i \times C}$)
$a_i(j), v_i(j)$	j-th feature at layer i ($\mathbb{R}^{C}$)
A_i, V_i	Feature resolution ($w_i^a \times h_i^a, w_i^v \times h_i^v$)
p_i^k	k-th learnable spatial embedding at layer i ($\mathbb{R}^{V_i \times C}, 0 \leq k < K$)
$\bar{p}_i$	Aligned feature at layer i ($\mathbb{R}^{V_i \times C}$)

3.1. Vision-to-Audio Knowledge Distillation

Cross-modal distillation from a visual teacher model to an audio model has two significant advantages: (i) training

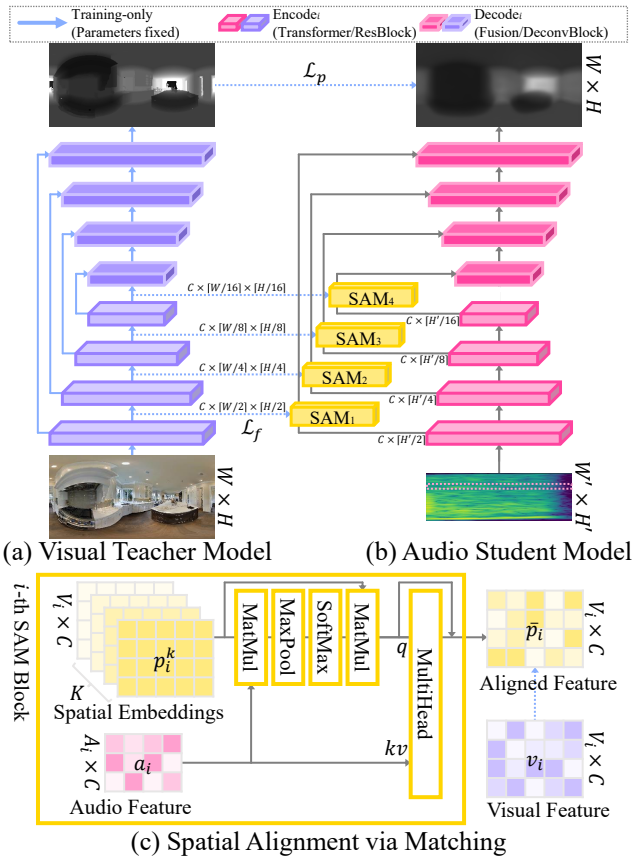


Figure 2: Overview of our spatial alignment via matching distillation framework.

without labeled data by turning to the teacher model's prediction (pseudo-GT) and (ii) teaching fine-grained knowledge to the student model via feature distillation. In general, cross-modal distillation for spatial reasoning leverages both pseudo-GT and feature outputs from one or more layers for fine-grained knowledge transfer [9]:

$$\mathcal{L}_{\text{crossKD}} = d(v_{\text{out}}, a_{\text{out}}) + \lambda \sum_i \sum_j d(v_i(j), a_i(j)), \quad (1)$$

where $d(\cdot, \cdot)$ is a distance function. This objective is well-defined for two modalities that are consistent up to pixel level (*e.g.*, distilling an RGB teacher to a depth student). On the other hand, it is less plausible to use the same method for vision-to-audio knowledge distillation.

The main difficulty that hinders knowledge transfer is the semantic and shape inconsistencies of the two heterogeneous modalities. First, the semantics of audio and visual features are not coherent with each other. For example, in the second term of Eq. (1), the j-th feature of an audio-only model at layer i may not always match the corresponding feature of a vision-only model. This lack of correspondence between the features of the two modalities makes direct distillation depicted in Fig. 1-(a) less effective, which is

empirically in line with previous research on vehicle tracking [5, 8]. Second, the shape of audio input is usually not identical to visual input, and simple interpolation of an audio input often deteriorates the performance. Moreover, it is even more challenging when the dimensions of the two modalities do not match, *e.g.*, predicting 3D surroundings from audio. Hence, it is necessary to establish a method that can effectively align with visual features regardless of specific input shapes other than naïve resizing or cropping.

3.2. Spatial Alignment via Matching

To resolve the challenges mentioned above, we introduce a novel method for cross-modal knowledge distillation of two heterogeneous modalities without semantic and shape consistency. We coin this method Spatial Alignment via Matching (SAM), which comprises three major components: input representation, learnable spatial embeddings, and feature refinement. To obtain the spatially aligned features for the i-th layer of the audio encoder, we can allocate a SAM block that accounts for both feature alignment and shape discrepancy, *i.e.*, $\text{SAM}_i : \mathbb{R}^{A_i \times C} \rightarrow \mathbb{R}^{V_i \times C}$.

Input Representation. Using Short-Term Fourier Transform (STFT) spectrograms of raw binaural audios, we can exploit any 2D deep networks as commonly done in audio representation learning [50, 51]. However, unlike previous works that rely on pseudo-GT [6, 49] or require identical shapes for feature-level distillation [5, 8], our method can be trivially applied where $(w_i^a, h_i^a) \neq (w_i^v, h_i^v)$.

In addition, SAM can handle more challenging scenarios like 1D encoders, *i.e.*, $w_i^a = 1$ or $h_i^a = 1$, by regarding the input spectrogram as a set of 1D patches. Decomposing the spectrogram into time bands ($W' \times 1$) or frequency bands ($1 \times H'$) can effectively reduce the feature shape and replace 2D with 1D operations. This allows for more efficient encoder implementation in terms of memory and time, making it applicable to memory-intensive scenarios.

Learnable Spatial Embeddings. It is essential to retain features that are spatially well-aligned with dense prediction output, especially when the input is not aligned with the output modality. In this regard, we design learnable spatial embeddings as a container to capture spatially varying information in paired audio-visual observations. We maintain a set of embeddings $p_i^0, ..., p_i^{K-1}$ identical in shape with visual features for each SAM and transform the shape of student features before the decoder. The number of learnable embeddings K may vary across layers, where more slots can be assigned to reconstruct high-level features.

For K learnable embeddings, we first derive a similarity matrix $T_i \in \mathbb{R}^{K \times V_i}$, which represents the proximity between provided audio feature a_i and the k-th spatial embedding. We compute the pairwise similarity between the j-th audio feature and the l-th feature in a spatial embedding, *i.e.*, $a_i(j), p_i^k(l) \in \mathbb{R}^C$, and select the maximum value along the j dimension:

$$T_i = \Big\|_{k=0}^{K-1} T_i^k = \Big\|_{k=0}^{K-1} \max_j p_i^k W_i a_i(j), \tag{2}$$

where $W_i \in \mathbb{R}^{C \times C}$ is a linear projection and $||$ is a concatenation operator. That is, higher similarity implies more coherency between the audio features and spatial embeddings at each region, allowing us to obtain features that are spatially aligned with the visual features.

By applying softmax along the K dimension of similarity matrix T_i, we then obtain a pooled embedding $\hat{p}_i \in \mathbb{R}^{V_i \times C}$ as a linear combination of embeddings:

$$\hat{p}_i = \Big\|_{l=0}^{V_i-1} \sum_{k=0}^{K-1} \frac{e^{T_i^k(l)}}{\sum_k e^{T_i^k(l)}} p_i^k(l). \tag{3}$$

The softmax term can be interpreted as a probability distribution of selecting k-th embedding for high audio-visual correspondence, making $\hat{p}_i$ coherent with audio features while maintaining the spatial structure of visual features.

Refinement with Student Features. For better coherence with audio features, we refine the pooled embedding $\hat{p}_i$ using audio feature a_i as keys and values by leveraging a multi-head attention mechanism (MultiHead) [52]:

$$\bar{p}_i = \text{MultiHead}(\hat{p}_i, a_i, a_i) + \hat{p}_i. \tag{4}$$

As a result, we obtain the aligned feature $\bar{p}_i$ from the SAM block at layer i. SAM can facilitate the spatial alignment between features at one (*i.e.*, a bottleneck between the encoder and decoder) or more layers. For instance, it can be applied to the global residual connection in pyramid-like architectures [11, 53, 54] to ensure shape consistency, as depicted in Fig. 2–(a-b).

3.3. Training and Inference

Network Architecture. For teacher models in each task, we follow the training procedure established in previous literature [12, 54, 13]. For simplicity, we train the teacher models using ground truth labels in the training split, while we also report the cross-modal distillation performance of non-iid settings in Appendix. We use ImageNet [55] pretrained weights for training teacher models in 2D tasks. Trained teacher models are only utilized during the training of a student model, with parameters fixed.

Our approach can be applied to a wide range of architectures for dense auditory prediction. We demonstrate this by using U-Net [11] with a ResNet-50 [56] backbone and Dense Prediction Transformers (DPT) [12] with a ViT-B/16 [57] backbone as representative examples of convolutional networks and vision transformer variants, respectively. We exploit Convolutional Occupancy Networks

(ConvONet) [13] as a base architecture for 3D reconstruction. Using paired audio-visual observations, student models are trained to mimic the output of the teacher model.

Learning Objective. We minimize the task-specific distance metric between the student and teacher model's prediction (pseudo-GT), *i.e.*, $\mathcal{L}_p = d(v_{\text{out}}, a_{\text{out}})$. To facilitate the cross-modal distillation, we integrate an auxiliary feature loss that promotes local coherence between a_i and v_i by optimizing the distance among triplets $(v_i(j), a_i(k), a_i(k'))$:

$$\mathcal{L}_f^i = \frac{1}{V_i} \sum_j \sum_{k' \in \mathcal{N}_k} \max(0, m - v_i(j) * a_i(k) + v_i(j) * a_i(k')), \tag{5}$$

where $m = 0.3$ is a margin, $\mathcal{N}_k$ is a set of negative samples regarding $a_i(k)$, and $*$ indicates cosine similarity. Since there are no ground truth positive pairs for local correspondence, we use $a_i(k) = \arg\max_{a_i(k)} a_i(k) * v_i(j)$ as a loosely defined positive pair. For $\mathcal{N}_k$, we either deem all the other features in a_i as negative or randomly select one among adjacent features, depending on the convergence of feature loss. In summary, our learning objective is as follows:

$$\mathcal{L}_{\text{Ours}} = \mathcal{L}_p + \lambda \sum_i \mathcal{L}_f^i, \tag{6}$$

where λ is a task-specific hyperparameter to balance the scale between the pseudo-GT loss and feature loss. We use up to four SAM blocks for all experiments, where we set $K = 64$ for the last SAM (SAM_4) and reduce the number by a factor of four. We train the student model from scratch, and during inference, we do not use any input, feature maps, or modules related to the visual modality; only the audio input and the trained audio-only student model are utilized. Further details are deferred to Appendix.

4. Experiments

We first discuss a new benchmark for three audio-based dense prediction tasks of scene understanding (§4.1). We then present the results of our approach for audio-based depth estimation, semantic segmentation, and 3D scene reconstruction tasks (§4.2–4.4).

4.1. The DAPS Benchmark

To evaluate the 2D and 3D dense prediction performance with audio, both the audio signal and the information regarding its surrounding space are required. Since none of the existing works benchmark multifaceted aspects of the omnidirectional surroundings as a whole, we organize a new benchmark upon existing simulators and datasets. We coin this benchmark Dense Auditory Prediction of Surroundings (DAPS). DAPS comprises 15.8K indoor scene observations with labels, where each sample consists of binaural audio, RGB panorama, and 3D voxel triples as observation and dense labels for three different tasks, as illustrated in Fig. 1-(b).

SoundSpaces [7] can simulate sound in indoor environments; for example, it includes Matterport3D [10] that deals with the material properties and layouts of a scene. Once setting the position and orientation of the recording agent in SoundSpaces, we obtain the recordings with respect to a set of emitter and receiver coordinate pairs. For simplicity, we report the results when the coordinates of an emitter and a receiver are identical.

After sampling coordinates information, we employ the Habitat simulator [58] to extract multimodal observations of a scene. We obtain RGB, depth, and semantic labels in equirectangular format from each location. To further collect 3D information of a scene, we extract the meshes surrounding the specified coordinate by truncating them, *i.e.*, 2.5m×2.5m×2m. Then, we use clustering-based filtering to remove noisy groups of meshes and keep only the most salient components. Finally, we generate 3D voxels from meshes for 3D reconstruction.

We carefully exclude the samples with weak auditory signals, such as outdoor scenes with high levels of noise, to maintain the quality of the benchmark. Specifically, for 2D dense prediction tasks, we eliminate samples whose labels have more than 10% missing pixels or noisy annotations. For 3D dense prediction, we exclude the samples with corrupted voxels by selecting the 95% lower confidence bound of the number of occupied voxels. We use 11.6K samples for training, 1.6K samples for validation, and 2.6K samples for testing in all experiments.

4.2. Results of Depth Estimation

4.2.1 Experiment Settings

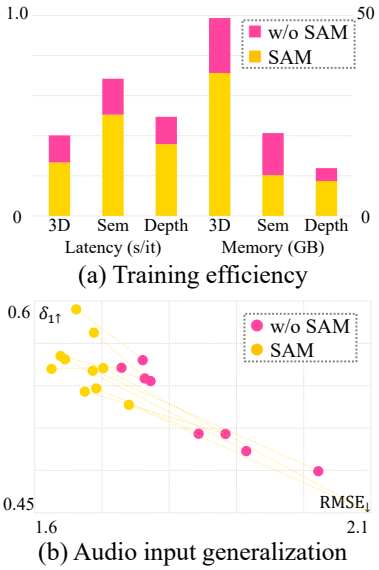
Following previous works on depth estimation [12, 59], we predict the depth of the whole surroundings given binaural audio from the scene. We follow the decoder design of [59] to train the model with the Inverse Huber loss. We report the results of sinusoidal sweep-convolved binaural inputs following the convention of [27, 28, 44, 46]. We also report the results of natural audio inputs [7] in Fig. 3-(b).

Evaluation Metrics. We report the mean absolute error (MAE), root mean squared error (RMSE), and delta accuracy $(\delta_1, \delta_2, \delta_3)$ for evaluation. MAE and RMSE reflect the error rate of our prediction, while the delta accuracy indicates the relative correctness of our prediction, *i.e.*, $\max(\frac{a_{\text{out}}}{v_{\text{out}}}, \frac{v_{\text{out}}}{a_{\text{out}}}) < 1.25^i$. To demonstrate the efficiency of our approach, we also report the memory allocation on GPU and latency during training.

Baselines. We include some state-of-the-art audio-only and distillation models as baselines [44, 8, 46], which are originally designed to predict bounding boxes or depth maps from a normal field-of-view with multi-channel audios. We also report the performance of losses proposed in

		MAE$_\downarrow$	RMSE$_\downarrow$	$\delta_{1\uparrow}$	$\delta_{2\uparrow}$	$\delta_{3\uparrow}$
	Teacher [11]	0.6524	1.1296	0.7633	0.8966	0.9328
	BilinearCoAttn [46]	1.2101	1.8366	0.5128	0.7009	0.8139
	BatVision [44]	0.9345	1.5740	0.6284	0.7975	0.8806
	MM-DistillNet [8]	0.8995	1.5812	0.6633	0.8178	0.8902
V→A U-Net [11]	Pseudo-GT ($\mathcal{L}_p$) [6]	0.9572	1.6436	0.6258	0.7971	0.8771
	+ Rank [5]	0.9524	1.6350	0.6279	0.7986	0.8786
	+ MTA [8]	0.9572	1.6392	0.6243	0.7956	0.8782
	+ SAM$_{\text{MultiHead}}$	0.8789	1.5604	0.6774	0.8256	0.8955
	+ SAM$_{\text{SpatialEmbeddings}}$	0.8760	1.5468	0.6787	0.8267	0.8965
	+ SAM$_{3,4(K=1)}$	0.8704	1.5467	0.6857	0.8302	0.8978
	+ **SAM$_{3,4}$**	**0.8633**	**1.5397**	**0.6869**	**0.8308**	**0.8982**
V→A DPT [12]	Pseudo-GT ($\mathcal{L}_p$) [6]	0.8926	1.5851	0.6684	0.8243	0.8943
	+ Rank [5]	0.9130	1.6017	0.6607	0.8159	0.8869
	+ MTA [8]	0.8913	1.5819	0.6694	0.8263	0.8953
	+ **SAM$_4$**	0.8517	**1.5276**	0.6971	0.8344	0.8986
	+ **SAM$_{3,4}$**	**0.8443**	1.5351	**0.7019**	**0.8392**	0.9000
	+ **SAM$_{1,2,3,4}$**	0.8497	1.5346	0.6992	0.8380	**0.9002**

Table 1: Comparison of depth estimation accuracy on DAPS-Depth test split.



(a) Training efficiency

(b) Audio input generalization

Figure 3: Analysis on distillation efficiency and input generalization.

	MAE$_\downarrow$	RMSE$_\downarrow$	$\delta_{1\uparrow}$
Mono	1.0783	1.7543	0.5829
16×16 Patch	0.8903	1.5786	0.6753
$1 \times H'$ Patch (freq.)	0.8902	1.5607	0.6629
$W' \times 1$ Patch (time)	**0.8497**	**1.5346**	**0.6992**
Embeddings$_{\text{NonSpatial}}$	0.8777	1.5334	0.6757
Embeddings$_{\text{Oracle}}$	0.5622	1.0308	0.8156

Table 2: Influence of input representation and learnable spatial embeddings in DPT+SAM on DAPS-Depth test split.

[6, 5, 8] combined with U-Net or DPT for fair comparison.

4.2.2 Results and Analyses

Comparison with Prior Arts. Table 1 summarizes the accuracy on DAPS-Depth test split. Compared to previous works on audio-only and distillation-based auditory depth estimation, our method achieves significant performance boosts across all metrics. For both U-Net and DPT, directly minimizing the feature distance between the teacher and the student (*i.e.*, +Rank/MTA) contributes marginally to the performance. Instead, adopting the proposed spatial alignment via matching improves the performance substantially, up to 10% (MAE) for U-Net. It is also worth noting that U-Net with SAM displays comparable performance with DPT variants. One of the important aspects of our approach is its efficiency, as illustrated in Fig. 3-(a). Compared to previous distillation methods, DPT+SAM improves both time and space efficiency by 27%, where the gap becomes wider for the other two tasks.

Ablation Studies. In Table 1, replacing full SAM blocks with multi-head attention (SAM$_{\text{MultiHead}}$) or learnable spatial embeddings (SAM$_{\text{SpatialEmbeddings}}$) deteriorates the absolute error rate by 1.5-1.8%. Reducing the number of spatial embeddings per layer to one (SAM$_{3,4(K=1)}$) is also harmful to performance. Increasing the number of SAM blocks for alignment can be beneficial, but forcefully matching the low-level vision features with audio features (*i.e.*, SAM$_{1,2}$) does not improve the prediction accuracy.

Table 2 analyzes the influence of different patch designs and spatial embeddings. Both frequency and time patches are more efficient than the regular patch, but only the time patch introduces significant performance gain. This implies that aggregating all frequency responses per short time span is a preferred input representation for dense auditory prediction. Also, the degraded performance of $\mathbb{R}^{K\times 1\times C}$ spatial embeddings instead of $\mathbb{R}^{K\times V_i\times C}$ (*i.e.*, non-spatial embeddings) stresses the importance of securing spatially varying information for matching. Finally, using actual visual features instead of learnable embeddings (*i.e.*, oracle embeddings) displays on-par performance with the teacher model.

Generalization to Natural Audio Inputs. Fig. 3-(b) reports the distillation performance of U-Net trained with diverse audio samples randomly selected from [7]. Not only our approach consistently achieves better performance, but the variance among different audio samples is also smaller than in previous distillation methods.

Qualitative Results. Fig. 4 displays the depth estimation results from binaural audio. Our approach can precisely measure the depth or structure of the room compared to prior arts. In some cases, it can even capture smaller objects like a billiards table in a scene from the audio.

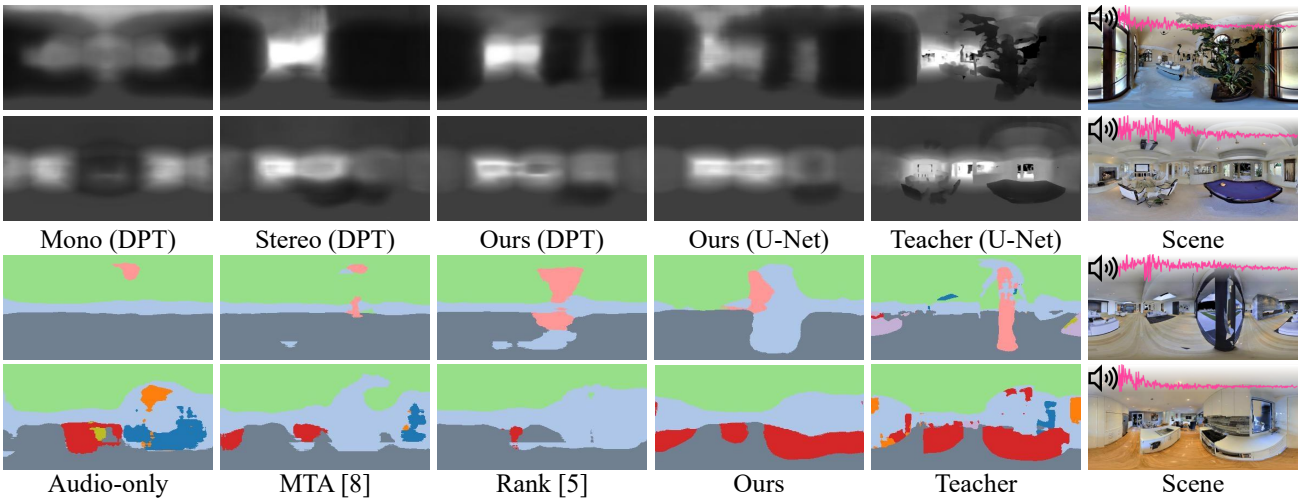


Figure 4: Qualitative examples of audio-based depth estimation (upper) and semantic segmentation (lower).

	pAcc$_\uparrow$	mAcc$_\uparrow$	mIoU$_\uparrow$	3IoU$_\uparrow$
Teacher	0.737	0.708	0.409	0.705
BilinearCoAttn [46]	0.605	0.493	0.340	0.538
MM-DistillNet [8]	0.629	0.515	0.311	0.581
Pseudo-GT ($\mathcal{L}_p$) [6]	0.628	0.513	0.320	0.576
+ MTA [8]	0.629	0.514	0.316	0.576
+ Rank [5]	0.642	0.520	0.359	0.587
+ **SAM$_{\text{Full}}$**	**0.644**	**0.526**	**0.363**	**0.600**

Table 3: Comparison of semantic segmentation accuracy on DAPS-Semantic test split.

4.3. Results of Semantic Segmentation

4.3.1 Experiment Settings

We train the audio student model to predict pixel-wise categories of the scene. Except for the pseudo-GT learning objective $\mathcal{L}_p$, we follow the training recipe explained in Sec. 4.2. As an auxiliary task, we predict the pseudo-GT segmentation with the penultimate layer feature for better performance, as proposed by Zhao *et al*. [54]. We train the model with the cross-entropy loss, where the primary and auxiliary loss ratio is 1:0.2.

Since it is virtually not tractable to classify 40+ semantic categories merely from the audio, we opt out classes about tiny objects (*e.g*., towels) and merge similar classes to establish nine classes for semantic segmentation based on audio. We report the performance of feature-level distillation methods with U-Net as a backbone.

Evaluation Metrics. We report the pixel-wise accuracy (pAcc), class-wise mean accuracy (mAcc), and class-wise mean IoU (mIoU) for all pixels with valid labels. Since it is challenging to label small objects in a scene with audio precisely, we introduce the mean IoU of ceiling, wall, and floor (3IoU) that constitutes a coarse layout of the scene.

4.3.2 Results and Analyses

Table 3 summarizes the semantic segmentation accuracy on DAPS-Semantic test split. Although predicting material properties or a semantic structure from auditory input is challenging, the result suggests that the overall output is acceptably plausible, achieving 87% of the teacher model's performance on the pAcc metric. Compared to depth estimation, the ranking-based objective fairly contributes to the distillation performance, which could be related to the classification error ensuring tighter bounds for ranking measures [60]. Still, SAM achieves better performance in all metrics, especially in predicting layout-relevant categories, *i.e*., +4% compared to Pseudo-GT.

Qualitative Examples. The last two rows of Fig. 4 illustrate the semantic segmentation results. Our approach can better predict the categories of smaller objects and the layout of the indoor surroundings, even under visually ill-posed scenarios like the windows in the third row.

4.4. Results of 3D Scene Reconstruction

4.4.1 Experiment Settings

We reconstruct a 3D scene with audio by means of voxel super-resolution. Voxel super-resolution aims to reconstruct high-resolution 3D objects using low-resolution voxelized meshes as input [61]. We use a teacher model that maps low (16^3) to high-resolution voxel grids (32^3) by capturing structural details of 3D shapes for reconstruction. Despite the difference in dimensions and shapes, the feature maps of the 3D teacher U-Net are utilized to learn the spatial alignment with auditory features, owing to the SAM blocks.

Evaluation Metrics. Following Peng *et al*. [13], we report IoU, Chamfer-L_1 distance, normal consistency (NC), and F1-score. We use IoU and F1 to measure the intersection between ground truths and predictions. Also, we

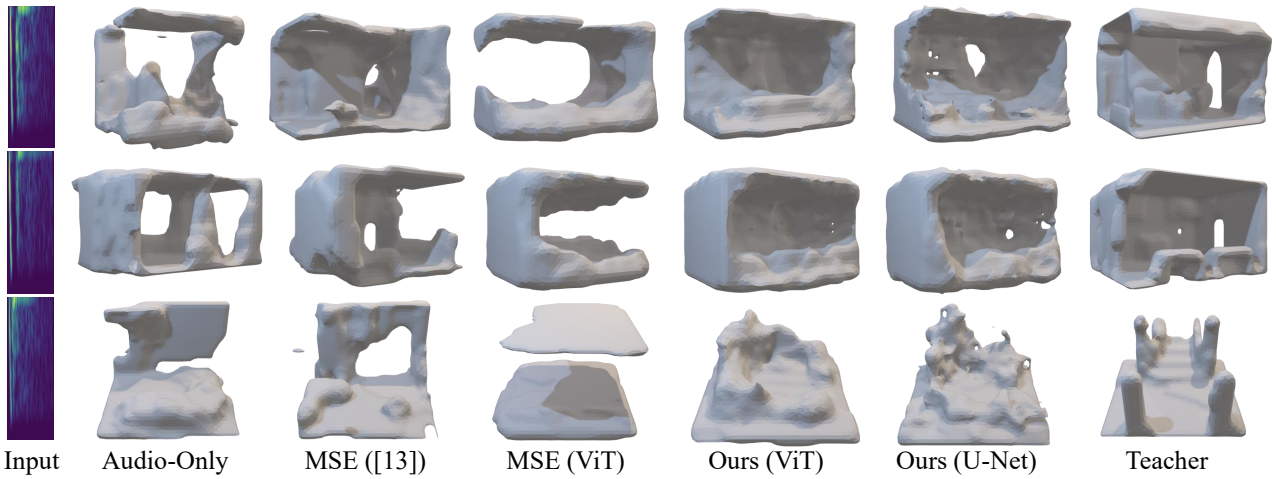


Figure 5: Qualitative examples of audio-based 3D scene reconstruction.

		IoU$_\uparrow$	Chamfer$_\downarrow$	NC$_\uparrow$	F1$_\uparrow$
Teacher [13]		0.548	0.0137	0.882	0.560
Audio-only$_{\text{Mono}}$		0.126	0.0698	0.625	0.189
Audio-only$_{\text{Stereo}}$		0.136	0.0643	0.639	0.196
[13]	MSE	0.137	0.0630	0.639	**0.203**
	Rank [5]	0.138	0.0636	0.640	0.200
	MTA [8]	0.149	0.0656	0.631	0.174
U-Net [11]	MSE	0.150	0.0676	0.626	0.177
	Rank [5]	0.153	0.0663	0.631	0.174
	MTA [8]	0.159	0.0660	0.645	0.170
	SAM$_{\text{Full}}$	**0.178**	**0.0555**	**0.679**	**0.203**
ViT [12]	MSE	0.154	0.0626	0.656	0.183
	Rank [5]	0.147	0.0698	0.671	0.177
	MTA [8]	0.154	0.0650	0.646	0.187
	SAM$_{\text{Full}}$	**0.178**	**0.0587**	**0.682**	**0.204**

Table 4: Comparison of 3D scene reconstruction accuracy on DAPS-3D test split.

evaluate Chamfer-L$_1$ distance and NC as similarity metrics based on multidimensional point sets and normal displacement vectors, respectively.

Baselines. Due to a lack of prior research on generating 3D objects from audio, we set up several conceivable baselines for comparison. First, we interpolate the 2D audio input to 3D to use ConvONet as a backbone. We report the performance of audio-only models and their variants with feature distillation. Second, as in the spatial alignment via matching framework, we use the 2D audio input as is and convert intermediate feature maps to match the shape of 3D features. We use U-Net [11] and ViT [57] as backbones to show that our approach can be applied to various encoder structures, where we include the ranking [5] or MTA [8] objectives for cross-modal distillation as baselines.

4.4.2 Results and Analyses

Table 4 reports the 3D scene reconstruction performance on DAPS-3D test split. Due to task difficulty, the performance gap between the teacher and the student is wider than 2D dense prediction tasks. Still, our approach improves the IoU score by 40% compared to audio-only models. Instead of forcefully converting the audio input representation, reducing the feature distance while keeping the audio input intact generally performs better. Lower Chamfer-L$_1$ scores of our approach, *i.e.*, an 18% reduction for the U-Net backbone, suggest that SAM facilitates the generation of points that are significantly closer to the ground truth.

Qualitative Examples. Fig. 5 visualizes our audio-based 3D scene reconstruction results. In the absence of visual cues, our approach accurately predicts the closed walls in a scene, even capturing details like holes (*e.g.*, doors or windows) and furniture. The substantial gap of quality between ours and prior arts in an open space (the last row of Fig. 5) stresses the importance of our distillation framework for dense prediction of 3D surroundings.

5. Conclusion

We addressed the audio-based dense prediction of indoor surroundings in 2D and 3D for the first time, addressing the challenges in vision-to-audio knowledge distillation: the discrepancy between the two modalities. To this end, we presented a novel spatial alignment via matching (SAM) distillation framework, accounting for local correspondence of multi-scale features with input shape inconsistency. In experiments in a newly collected DAPS dataset, our distillation framework consistently improves the performance across multiple tasks ranging from 2D to 3D with various architectures as backbones. Qualitative results indicate that our approach better captures fine-grained information about the scene from the auditory input compared to prior arts.

Acknowledgement. This work was supported by LG AI Research, National Research Foundation of Korea (NRF) grant (No.2023R1A2C2005573) and Institute of Information & Communications Technology Planning & Evaluation (IITP) grant (No.2022-0-00156, 2019-0-01082, 2021-0-01343) funded by the Korea government (MSIT). Gunhee Kim is the corresponding author.

References

[1] John William Strutt Baron Rayleigh. *The theory of sound*. 1896. 1

[2] E Colin Cherry. Some experiments on the recognition of speech, with one and with two ears. *The Journal of the acoustical society of America*, 1953. 1

[3] David V Smith, Ben Davis, Kathy Niu, Eric W Healy, Leonardo Bonilha, Julius Fridriksson, Paul S Morgan, and Chris Rorden. Spatial attention evokes similar activation patterns for visual and auditory stimuli. *Journal of cognitive neuroscience*, 2010. 1

[4] Giorgia Cona and Cristina Scarpazza. Where is the "where" in the brain? a meta-analysis of neuroimaging studies on spatial cognition. *Human brain mapping*, 2019. 1

[5] Chuang Gan, Hang Zhao, Peihao Chen, David Cox, and Antonio Torralba. Self-supervised moving vehicle tracking with stereo sound. In *CVPR*, 2019. 1, 2, 3, 4, 6, 7, 8

[6] Arun Balajee Vasudevan, Dengxin Dai, and Luc Van Gool. Semantic object prediction and spatial sound super-resolution with binaural sounds. In *ECCV*, 2020. 1, 3, 4, 6, 7

[7] 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. 1, 2, 5, 6

[8] 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. 1, 2, 3, 4, 5, 6, 7, 8

[9] Saurabh Gupta, Judy Hoffman, and Jitendra Malik. Cross modal distillation for supervision transfer. In *CVPR*, 2016. 1, 2, 3

[10] 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. 2, 5

[11] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015. 2, 4, 6, 8

[12] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *ICCV*, 2021. 2, 4, 5, 6, 8

[13] Songyou Peng, Michael Niemeyer, Lars M. Mescheder, Marc Pollefeys, and Andreas Geiger. Convolutional occupancy networks. In *ECCV*, 2020. 2, 4, 5, 7, 8

[14] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. *ECCV*, 2012. 2

[15] Iro Armeni, Sasha Sax, Amir R Zamir, and Silvio Savarese. Joint 2d-3d-semantic data for indoor scene understanding. *arXiv:1702.01105*, 2017. 2

[16] Iro Armeni, Ozan Sener, Amir R Zamir, Helen Jiang, Ioannis Brilakis, Martin Fischer, and Silvio Savarese. 3d semantic parsing of large-scale indoor spaces. In *CVPR*, 2016. 2

[17] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *CVPR*, 2017. 2

[18] Camille Couprie, Clément Farabet, Laurent Najman, and Yann LeCun. Indoor semantic segmentation using depth information. In *ICLR*, 2013. 2

[19] Feihu Zhang, Jin Fang, Benjamin Wah, and Philip Torr. Deep fusionnet for point cloud semantic segmentation. In *ECCV*, 2020. 2

[20] Huy Ha and Shuran Song. Semantic abstraction: Open-world 3d scene understanding from 2d vision-language models. In *CoRL*, 2022. 2

[21] Junha Roh, Karthik Desingh, Ali Farhadi, and Dieter Fox. Languagerefer: Spatial-language model for 3d visual grounding. In *CoRL*, 2022. 2

[22] Ruohan Gao and Kristen Grauman. 2.5 d visual sound. In *CVPR*, 2019. 2

[23] Changan Chen, Ruohan Gao, Paul Calamia, and Kristen Grauman. Visual acoustic matching. In *CVPR*, 2022. 2

[24] Nikhil Singh, Jeff Mentch, Jerry Ng, Matthew Beveridge, and Iddo Drori. Image2reverb: Cross-modal reverb impulse response synthesis. In *ICCV*, 2021. 2

[25] Changan Chen, Carl Schissler, Sanchit Garg, Philip Kobernik, Alexander Clegg, Paul Calamia, Dhruv Batra, Philip W Robinson, and Kristen Grauman. Soundspaces 2.0: A simulation platform for visual-acoustic learning. In *NeurIPS Datasets and Benchmarks Track*, 2022. 2

[26] Senthil Purushwalkam, Sebastia Vicenc Amengual Gari, Vamsi Krishna Ithapu, Carl Schissler, Philip Robinson, Abhinav Gupta, and Kristen Grauman. Audio-visual floorplan reconstruction. In *ICCV*, 2021. 2

[27] Ruohan Gao, Changan Chen, Ziad Al-Halah, Carl Schissler, and Kristen Grauman. Visualechoes: Spatial image representation learning through echolocation. In *ECCV*, 2020. 2, 5

[28] Kranti Kumar Parida, Siddharth Srivastava, and Gaurav Sharma. Beyond image to depth: Improving depth prediction using echoes. In *CVPR*, 2021. 2, 5

[29] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. In *NIPS Deep Learning and Representation Learning Workshop*, 2014. 2

[30] Judy Hoffman, Saurabh Gupta, Jian Leong, Sergio Guadarrama, and Trevor Darrell. Cross-modal adaptation for rgb-d 'detection. In *ICRA*, 2016. 2

[31] Nuno C Garcia, Pietro Morerio, and Vittorio Murino. Modality distillation with multiple stream networks for action recognition. In *ECCV*, 2018. 2

[32] Jianbo Jiao, Yunchao Wei, Zequn Jie, Honghui Shi, Rynson WH Lau, and Thomas S Huang. Geometry-aware distillation for indoor semantic segmentation. In *CVPR*, 2019. 2

[33] Antonin Vobecky, David Hurych, Oriane Siméoni, Spyros Gidaris, Andrei Bursuc, Patrick Pérez, and Josef Sivic. Drive&segment: Unsupervised semantic segmentation of urban scenes via cross-modal distillation. In *ECCV*, 2022. 2

[34] Mingmin Zhao, Tianhong Li, Mohammad Abu Alsheikh, Yonglong Tian, Hang Zhao, Antonio Torralba, and Dina Katabi. Through-wall human pose estimation using radio signals. In *CVPR*, 2018. 2

[35] Siddharth Roheda, Benjamin S Riggan, Hamid Krim, and Liyi Dai. Cross-modality distillation: A case for conditional generative adversarial networks. In *ICASSP*, 2018. 2

[36] Yanbei Chen, Yongqin Xian, A Koepke, Ying Shan, and Zeynep Akata. Distilling audio-visual knowledge by compositional contrastive learning. In *CVPR*, 2021. 2

[37] Ruohan Gao, Tae-Hyun Oh, Kristen Grauman, and Lorenzo Torresani. Listen to look: Action recognition by previewing audio. In *CVPR*, 2020. 2

[38] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. Asr is all you need: Cross-modal distillation for lip reading. In *ICASSP*, 2020. 2

[39] Ya Zhao, Rui Xu, Xinchao Wang, Peng Hou, Haihong Tang, and Mingli Song. Hearing lips: Improving lip reading by distilling speech recognizers. In *AAAI*, 2020. 2

[40] Yanpeng Zhao, Jack Hessel, Youngjae Yu, Ximing Lu, Rowan Zellers, and Yejin Choi. Connecting the dots between audio and text without parallel data through visual knowledge transfer. In *NAACL*, 2022. 2

[41] Xufeng Zhao, Cornelius Weber, Muhammad Burhan Hafez, and Stefan Wermter. Impact makes a sound and sound makes an impact: Sound guides representations and explorations. In *IROS*, 2022. 2

[42] Masahiro Yasuda, Yasunori Ohishi, and Shoichiro Saito. Echo-aware adaptation of sound event localization and detection in unknown environments. In *ICASSP*, 2022. 2

[43] Cho-Ying Wu, Chin-Cheng Hsu, and Ulrich Neumann. Cross-modal perceptionist: Can face geometry be gleaned from voices? In *CVPR*, 2022. 2

[44] Jesper Haahr Christensen, Sascha Hornauer, and X Yu Stella. Batvision: Learning to see 3d spatial layout with two ears. In *ICRA*, 2020. 3, 5, 6

[45] Ethan Tracy and Navinda Kottege. Catchatter: Acoustic perception for mobile robots. *IEEE RA-L*, 2021. 3

[46] Go Irie, Takashi Shibata, and Akisato Kimura. Co-attention-guided bilinear model for echo-based depth estimation. In *ICASSP*, 2022. 3, 5, 6, 7

[47] Ziyang Chen, Xixi Hu, and Andrew Owens. Structure from silence: Learning scene structure from ambient sound. In *CoRL*, 2022. 3

[48] Alexander Raistrick, Nilesh Kulkarni, and David F Fouhey. Collision replay: What does bumping into things tell you about scene geometry? In *BMVC*, 2021. 3

[49] Dengxin Dai, Arun Balajee Vasudevan, Jiri Matas, and Luc Van Gool. Binaural soundnet: predicting semantics, depth and motion with binaural sounds. *IEEE TPAMI*, 2022. 3, 4

[50] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. Cnn architectures for large-scale audio classification. In *ICASSP*, 2017. 4

[51] Yuan Gong, Yu-An Chung, and James Glass. Ast: Audio spectrogram transformer. *arXiv:2104.01778*, 2021. 4

[52] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NIPS*, 2017. 4

[53] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *CVPR*, 2017. 4

[54] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *CVPR*, 2017. 4, 7

[55] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 4

[56] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 4

[57] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 4, 8

[58] Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, Devi Parikh, and Dhruv Batra. Habitat: A Platform for Embodied AI Research. In *ICCV*, 2019. 5

[59] Hualie Jiang, Zhe Sheng, Siyu Zhu, Zilong Dong, and Rui Huang. Unifuse: Unidirectional fusion for 360 panorama depth estimation. *IEEE RA-L*, 2021. 5

[60] Wei Chen, Tie-Yan Liu, Yanyan Lan, Zhi-Ming Ma, and Hang Li. Ranking measures and loss functions in learning to rank. *NIPS*, 2009. 7

[61] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *CVPR*, 2019. 7