

Disentangled Acoustic Fields For Multimodal Physical Scene Understanding

Yin, Jie; Luo, Andrew; Du, Yilun; Cherian, Anoop; Marks, Tim K.; Le Roux, Jonathan; Gan, Chuang

TR2024-125 September 07, 2024

Abstract

We study the problem of multimodal physical scene understanding, where an embodied agent needs to find fallen objects by inferring object properties, direction, and distance of an impact sound source. Previous works adopt feed-forward neural networks to directly regress the variables from sound, leading to poor generalization and domain adaptation issues. In this paper, we illustrate that learning a disentangled model of acoustic formation, referred to as disentangled acoustic field (DAF), to capture the sound generation and propagation process, enables the embodied agent to construct a spatial uncertainty map over where the objects may have fallen. We demonstrate that our analysis-by-synthesis framework can jointly infer sound properties by explicitly decomposing and factorizing the latent space of the disentangled model. We further show that the spatial uncertainty map can significantly improve the success rate for the localization of fallen objects by proposing multiple plausible exploration locations.

IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) 2024

Disentangled Acoustic Fields For Multimodal Physical Scene Understanding

Jie Yin¹, Andrew Luo², Yilun Du³, Anoop Cherian⁴, Tim K. Marks⁴, Jonathan Le Roux⁴, Chuang Gan^{5*}

Abstract—We study the problem of multimodal physical scene understanding, where an embodied agent needs to find fallen objects by inferring object properties, direction, and distance of an impact sound source. Previous works adopt feed-forward neural networks to directly regress the variables from sound, leading to poor generalization and domain adaptation issues. In this paper, we illustrate that learning a disentangled model of acoustic formation, referred to as disentangled acoustic field (DAF), to capture the sound generation and propagation process, enables the embodied agent to construct a spatial uncertainty map over where the objects may have fallen. We demonstrate that our analysis-by-synthesis framework can jointly infer sound properties by explicitly decomposing and factorizing the latent space of the disentangled model. We further show that the spatial uncertainty map can significantly improve the success rate for the localization of fallen objects by proposing multiple plausible exploration locations.

I. INTRODUCTION

Imagine walking through a forest with your eyes closed, listening to the sounds around you. As you move, you hear the rustling of leaves as an animal passes by, the gentle bubbling of a nearby stream, and the soft whisper of the wind. These sounds provide valuable information about the environment. Sound waves are influenced by the objects they encounter, changing in timbre, direction, and intensity as they reflect, diffract, and absorb. As humans, we intuitively understand how sound behaves in physical spaces, enabling us to infer the presence, location, and physical properties of objects from the sounds we hear.

Recent progress in neural fields has yielded high-fidelity models of perceptual modalities such as vision, touch, and sound. Most recently, neural acoustic fields (NAFs) [1] propose representing spatial acoustics of sounds, enabling continuous modeling of sound propagation and reverberation in a given scene. By modeling such acoustics, NAFs implicitly capture the structure and material properties of a scene. However, NAFs are overfitted to the acoustic properties of a single room, preventing them from being used as a disentangled model of sound across many environments.

In this work, we propose disentangled acoustic fields (DAFs), an approach to modeling acoustic properties across a multitude of different scenes. In NAFs, the short-time Fourier transform (STFT) of audio reverberation is the object of the modeling, but it is highly sensitive to the geometry of each

scene and thus difficult to fit across different scenes. Instead, DAFs seek to model object sounds across multiple scenes using the power spectral density (PSD). This approach provides a lower dimensional compact representation of acoustics that preserves much of the physical information in emitted sounds. We demonstrate the effectiveness of this approach by showing high-accuracy object property inference across a set of different scenes.

We demonstrate how DAFs can be used to effectively enhance audio perception. Specifically, we propose using DAFs as a “mental simulation engine” that can test different physical scene configurations to identify the world state that best matches the given sound. This “analysis-by-synthesis” approach allows us to robustly infer the underlying locations of fallen objects and effectively navigate to locate them. Our experiments show that, by using DAFs, we can accurately identify categories of fallen objects and their locations, even in complex acoustic environments.

Acoustic rendering with DAFs further enables us to obtain an uncertainty measure of different physical scene parameters, such as object locations, by assessing the mismatch between a simulated sound and ground-truth sound. We illustrate how such uncertainty may be used in the task of finding a fallen object, where we may naturally generate plans to different goals by considering the underlying uncertainty cost. In summary, our contributions are as follows:

- We introduce Disentangled Acoustic Fields (DAFs), an approach to model acoustic properties across a multitude of different scenes.
- We illustrate how analysis-by-synthesis using DAFs enables us to infer the physical properties of a scene.
- We illustrate how we may use DAFs to represent uncertainty and to navigate and find fallen objects.

II. RELATED WORK

A. Neural Implicit Representations

Learned implicit functions have emerged as a promising representation of the 3D geometry [2], [3], [4], appearance [5], [6], and acoustics of a scene [1]. Unlike traditional discrete representations, implicit functions compactly encode information in the weights of a neural network, and can continuously map from spatial coordinates to output. Recent work has proposed to encode shapes as signed distance fields, learn appearance with differentiable rendering, and render acoustics by generating spectrograms [7] [1]. For acoustics, [8] proposed to jointly generate acoustics and images by sampling from a joint manifold, and [1] introduced the concept of Neural Acoustic Fields (NAFs), an implicit representation

¹ is with Shanghai Jiao Tong University, ² is with Carnegie Mellon University, ³ is with Massachusetts Institute of Technology, ⁴ is with Mitsubishi Electric Research Laboratories(MERL), ⁵ is with MIT-IBM AI Lab and UMass Amherst. * Corresponding Author: Chuang Gan (ganchuang@csail.mit.edu).

This research was supported by MERL.

that captures sound propagation in a physical scene. While NAFs enable the modeling of sounds at novel locations in a single scene, they cannot be generalized to enable reasoning across novel scenes. In contrast, our method can generalize to novel scenes at test time, enables joint inference of object properties and location, and allows uncertainty-aware object localization.

B. Multimodal Scene Understanding

Recent work has explored the use of input modalities beyond vision alone for scene understanding [9], [10]. Extensive studies have demonstrated the effectiveness of integrating audio and visual information in diverse scene understanding applications [11], [12], [13], [14]. For instance, [15], [16], [17] employ visual input to separate and localize sounds, [18] leverages spatial cues contained in echoes for more accurate depth estimation, while [19], [1], [20], [21] demonstrate the potential of sound in learning multimodal features and inferring scene structure. Cross-modal generation has gained increasing attention by researchers [22], [23], [24], [25]. Furthermore, [26], [27], [28] integrate both visual and auditory information to localize target objects more accurately. Motivated by these findings, we propose a disentangled acoustic field for physical scene understanding, where an embodied agent seeks to find fallen objects by inferring their physical properties, direction, and distance from an impact sound.

C. Audio-Visual Navigation

Our work is also closely related to audio-visual navigation, where navigation is achieved using audio signals to augment vision [29], [30], [31]. In particular, [32] proposed the AudioGoal challenge, where an embodied agent is required to navigate to a target emitting a constant sound using audio for positional cues [33]. Building on this, [29] introduced the novel task of semantic audio-visual navigation, in which the agent must navigate to an object with semantic vision and short bursts of sound. However, their dataset had one limitation: it did not include synthesis capability for impact sound and thus could not render the physical properties (like material, position) by audio. To address both issues, [34] proposed the Find Fallen Object task, where physical reasoning was combined with sound. This dataset was based on the TDW [35] simulation platform, which can generate audio from a variety of materials and parameters, and utilizes Resonance Audio [36] (a technology for accurately replicating how sound interacts with the environment in 3D spaces) to spatialize the impact sounds depending on the room’s spatial dimensions and wall/floor materials. Considering these advantages, we choose it as the benchmark to assess the capability of our proposed method on multi-modal physical scene understanding.

III. PROPOSED METHOD

We are interested in learning a disentangled framework of sound that can effectively generalize across scenes, object types, and object locations. Key to our approach is an

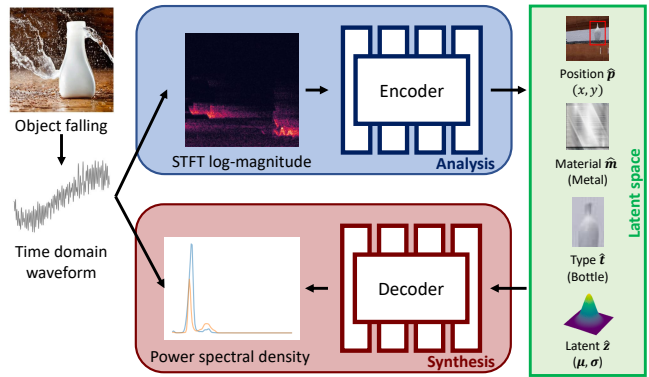


Fig. 1. **Illustration of DAFs.** The encoder maps the binaural short-time Fourier transform (STFT) of the audio input into a new space containing physical audio information such as object position, material, type, and a continuous latent. The decoder utilizes these parameters to reconstruct the power spectral density (PSD) of the audio. The two components form an analysis-by-synthesis loop capable of inferring object properties, and are jointly learned during training.

explicitly decomposed latent space that models the individual contribution of the sound factors. We first describe the parameterization of our disentangled acoustic field, which simultaneously enables factorization of the sound generation process and is defined on continuous locations in space. We further characterize the design choices that enable robust generalization and describe how we can use the continuous nature of our disentangled acoustic field to facilitate the localization of a fallen object.

A. Physics of Sound

Given the sound of a falling object received by an agent as binaural signal s , we seek to identify the relative egocentric object location $p \in \mathbb{R}^3$, the object material category $m \in \{1, 2, \dots, M\}$, the object type category $t \in \{1, 2, \dots, T\}$, and a low-dimensional latent code $z \in \mathcal{R}^k$, where z can contain information that is independent from the previous factors, such as information about the scene structure, scene materials, and properties about the propagation medium. Given an accurate model of sound formation G , we seek to reconstruct the sound via $G(p, m, t, z)$. In practice, we here do not reconstruct the sound itself, but its power spectral density, a simplified representation encompassing essential properties about the falling object.

B. Disentangled Acoustic Fields (DAFs)

We aim to learn a disentangled model of sound formation that facilitates efficient inference for characterizing sound properties. The parameterization of sound formation introduced in Section III-A provides a general framework for building such a disentangled model. To enhance effective learning, we structure our framework using an encoder, denoted as E_ω , and a generator, denoted as G_ϕ . By instructing the network to consider the relative egocentric location of the sound emitter, we guide it to disregard the absolute positions of the agent and object, and reason in a generalizable fashion.

Given a sound signal represented as a binaural waveform $s \in \mathbb{R}^{2 \times t}$, we process the signal using the short-time Fourier transform (STFT), and retain the log-magnitude as S . Following prior work, we discard the phase present in the original signal, which is difficult to model [8], [1]. We further investigated the choice of output representation, and found that the STFT of a fallen object sound used in prior work [34] includes large irregular stretches of uninformative silence along the time domain. The irregular and unpredictable temporal gaps are difficult for a neural network to effectively estimate and reconstruct, and in practice, a full-fidelity reconstruction of the sound signal in the original domain may not be necessary, as our ultimate goal is the inference of the underlying factors of sound formation. We thus transform the sound from the waveform domain into power spectral density (PSD) representation \bar{S} using Welch’s method (basically an average pooling over time of the squared STFT magnitude) as a target for our generator, which retains crucial information on the power present in each frequency, but collapses the information along the time dimension.

We may thus model the latent factors as the outputs of an encoder E_ω which takes as input the sound representation S :

$$(\hat{p}, \hat{m}, \hat{t}, \hat{\mu}, \hat{\sigma}) = E_\omega(S); \quad \hat{z} \sim \mathcal{N}(\hat{\mu}, \hat{\sigma}^2 \cdot \mathbb{I}), \quad (1)$$

where we model \hat{z} as a sample from a diagonal Gaussian with mean $\hat{\mu}$ and standard deviation $\hat{\sigma}$. This restricted parameterization prevents the encoder from compressing all sound information into \hat{z} . The generator G_ϕ is modeled as a neural field which takes as input the latent factors and attempts to generate the PSD:

$$\hat{S} = G_\phi(\hat{p}, \hat{m}, \hat{t}, \hat{z}). \quad (2)$$

We train our network with supervision on the latent factors and the output. For the i -th training example, we have access to the ground truth location, object material, and object type as the tuple (p_i, m_i, t_i) .

The object type and material are supervised with cross-entropy loss:

$$\mathcal{L}_{\text{type}} = \text{CrossEntropy}(t_i, \hat{t}_i), \quad (3)$$

$$\mathcal{L}_{\text{material}} = \text{CrossEntropy}(m_i, \hat{m}_i), \quad (4)$$

where t_i and m_i are the ground-truth object type and material for the i -th training sample, and \hat{t}_i and \hat{m}_i their estimates. An MSE loss is applied to facilitate the learning of the position vector:

$$\mathcal{L}_{\text{position}} = \frac{1}{2} \sum_i \|\hat{p}_i - p_i\|_2^2. \quad (5)$$

During training, we sample from the posterior for a given sound S_i modeled as a multivariate Gaussian with diagonal covariance:

$$q_\omega(z|S_i) := \mathcal{N}(z; \mu_i, \sigma_i^2 \cdot \mathbb{I}) \quad (6)$$

We apply the reparameterization trick [37] to allow for backpropagation through the parameters μ_i, σ_i , by setting

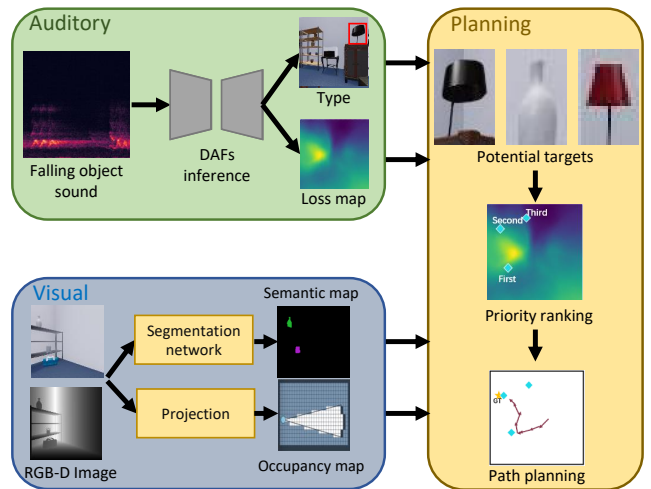


Fig. 2. **Planning with DAFs.** The agent jointly uses auditory and visual information as part of the planning process. The auditory branch takes as input the sound S represented as STFT. Using the DAF, we infer the factors responsible for the sound production including possible object types and a reconstruction loss map for each potential object location. The visual branch takes as input RGB-D images and provides a semantic map and occupancy map to the planner. The planner combines the information and uses the loss map to produce a priority list of locations. Path planning is completed using the A^* algorithm.

$z_i = \mu_i + \sigma_i \odot \epsilon$, where $\epsilon \sim \mathcal{N}(0, \mathbb{I})$. The latent z is regularized with:

$$D_{\text{KL}}(q_\omega(z|S_i) || \mathcal{N}(0, \mathbb{I})) \quad (7)$$

The output of the generator is supervised with an MSE loss to facilitate the prediction of the PSD:

$$\mathcal{L}_{\text{PSD}} = \frac{1}{d} \sum_i \|\hat{S}_i - \bar{S}_i\|^2, \quad (8)$$

where d is the dimension of the output PSD feature. In summary, our overall objective is to minimize:

$$\mathcal{L}_{\text{total}} = \alpha \mathcal{L}_{\text{type}} + \beta \mathcal{L}_{\text{material}} + \gamma \mathcal{L}_{\text{position}} + \delta D_{\text{KL}} + \eta \mathcal{L}_{\text{PSD}}, \quad (9)$$

where $(\alpha, \beta, \gamma, \delta, \eta)$ are hyperparameters.

C. Inference of Sound Parameters

We seek to invert the model of sound generation and compute the factors responsible for a given sound. Our disentangled model enables us to use “analysis by synthesis” to render all different acoustics parameters and find the one which matches our sound the best. However, the combinatorial complexity of enumerating combinations of factors renders this process computationally prohibitive. Instead, we amortize the inference of type, material, and latent into the joint learning process of the encoder and generator. We focus our efforts on the inference of object position, which is composed of continuous variables and is critical for the localization of the fallen object.

Loss Map Generation: Given a sound s as recorded by an embodied agent, we use the encoder to infer the material, type, and continuous latent. We define a search space 10 m

$\times 10$ m centered on the agent position, and discretize this space using a resolution of 0.1 m. Using the previously inferred type, material, and Gaussian latent, we iterate over all possible locations p_j where the object could be located. The current iterated position is combined with the other factors as inferred by the encoder network and provided to the generator. The pipeline is shown in Algorithm 1.

The generated PSD is compared against the ground-truth PSD \bar{S} of the sound of the fallen object, and an MSE difference is recorded at each location. In this fashion, we may generate a loss map corresponding to the two-dimensional coordinates. Since the loss map is based on the egocentric neural acoustic field, we need to convert the relative coordinates to the global frame of the room for path planning:

$$f_{r2g,c,\theta} \left(\begin{bmatrix} x \\ y \end{bmatrix} \right) = \begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} c_x \\ c_y \end{bmatrix} \quad (10)$$

where (c_x, c_y) is the agent’s current position in the global frame and θ is its rotation angle in the global frame, while (x, y) is the coordinate in the agent’s egocentric frame.

Uncertainty-Aware Planning: We adopt a modular approach to path planning [34]. Given the audio, we predict the object type and location via E_ω . We further construct an internal model of the world from RGB-D images of the environment. Semantic segmentation is computed using a pre-trained Mask-RCNN [38] model as illustrated in Figure 3. We use the depth map to project the semantic map into world space. Given the object types as inferred by the encoder network, we take the top-3 object type candidates (local minimas) as search targets. If a target is visible to the agent when initialized, the agent attempts to navigate to the location of the visible object. If there is no target in view, the agent will navigate to the predicted position. If the first attempt fails, the agent updates the world model and searches for potential object candidates. Object candidates are ranked according to the loss map value at the location corresponding to each object. Once the target list is determined, we apply the A^* algorithm [39] to plan the shortest collision-free path to the first target in an unvisited area of the map.

Algorithm 1: Inferring a Loss Map of Positions

Input: Sound as log-magnitude STFT S and PSD \bar{S} , encoder network E_ω , generator network G_ϕ , loss map grid $\mathcal{L}_{\text{grid}}$, function $f_{r2g,c,\theta}$ for global coordinates

- 1: $(\hat{p}, \hat{m}, \hat{t}, \hat{z}) = E_\omega(S)$
 - 2: for $x_{\text{pos}}, y_{\text{pos}}$ in $[-5 \text{ m}, +5 \text{ m}]$:
 - 3: $\hat{p} = (x_{\text{pos}}, y_{\text{pos}})$
 - 4: $\mathcal{L}_{\text{grid}}[f_{r2g,c,\theta}(\hat{p})] = \|G_\phi([\hat{p}, \hat{m}, \hat{t}, \hat{z}]) - \bar{S}\|_2^2$
-

IV. EXPERIMENT

A. Inference of Object Properties

To test the physical inference capability of our proposed model, we first evaluate it on the Find Fallen Dataset [34]

and compare it against two baselines. The first is the modular sound predictor presented in [34], which was learned without the use of a disentangled model. The second is a gradient-based optimization process that minimizes the difference between the predicted and ground-truth PSD by optimizing all latent factors using our trained generator. All methods are evaluated on the same test split. To enable exploring multiple plausible locations, we mark a type as accurately predicted if the correct object type is within the top-3 predicted categories. The results in Table II show that our model significantly outperforms the baseline methods in both position and type prediction accuracy. By jointly learning a disentangled model along the encoder, we can more accurately predict the object location and the object type. Gradient-based optimization fails in jointly optimizing the position and object type, and is easily stuck in local minima.

We evaluate our model on the OBJECTFOLDER2 [40] dataset, which contains 1000 virtualized objects with different sizes and types along with acoustic, visual, and tactile sensory information. Correspondingly, we adjust the output of the encoder network and the input of the decoder network to correspond to the object scale and type. As indicated in Table I, our method surpasses the performance of the baseline approach [40], which employs a ResNet-18 architecture utilizing the magnitude spectrogram as input to predict the object scale and type.

TABLE I
COMPARISON OF SCALE ESTIMATION AND TYPE PREDICTION
ACCURACY ON OBJECTFOLDER2.

Method	Scale (m) ↓	Type Acc. ↑
Baseline [40]	0.20	0.98
Ours	0.17	0.99

We further extend our evaluation to the REALIMPACT dataset [41], encompassing 150,000 impact sound recordings across 50 **real-world** object categories. This dataset incorporates impact sounds captured from 600 distinct listener locations, spanning 10 angles, 15 heights, and 4 distances. We use the same train/test split across all methods. To accommodate with the dataset, we adapt the output of the encoder network and the input of the generator network to be angle, height, and distance. The official baseline method is a ResNet-18 network employing the magnitude spectrogram of impact sounds to predict the corresponding object properties. As highlighted in Table III, our method demonstrates a significant improvement over baseline methods in predicting the category of object angle, height, and distance.

B. Navigation and Planning

Experimental Setup. We use the TDW [35] simulation platform to evaluate our proposed method. Our networks are trained on the Find Fallen Object dataset¹, following the experimental configuration described in [34]. This dataset contains 8000 instances of 30 physical object types in 64

¹https://github.com/chuanggf/find_fallen_objects

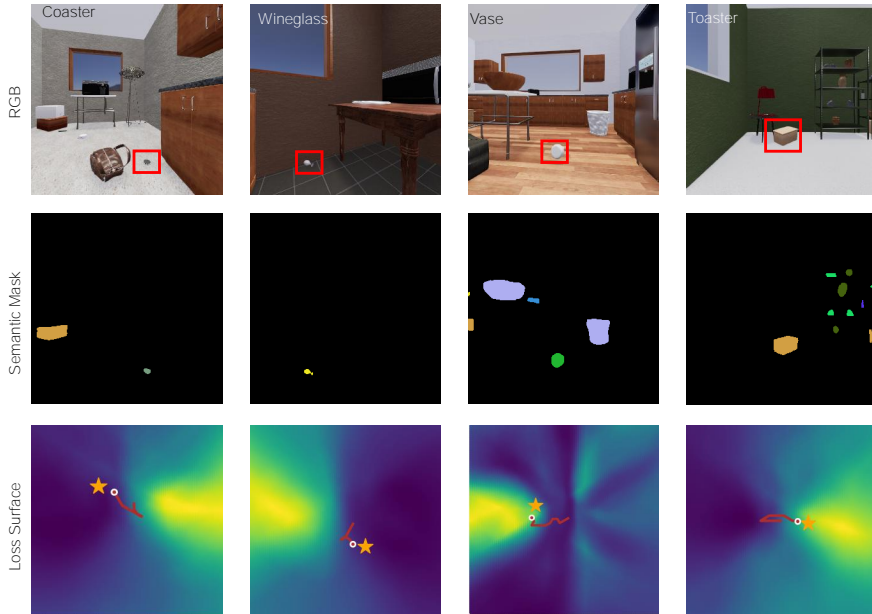


Fig. 3. **Visualization of visual input and the sound-derived loss map in four scenes.** **Top:** RGB images of the agent’s view with the target object in a red bounding box. **Middle:** Semantic map produced from the RGB images. **Bottom:** The red line indicates the path the agent takes, with the end point shown as a circular dot. The ground-truth object location is shown as a gold star.

TABLE II
COMPARISON OF POSITION ERROR AND TYPE PREDICTION
ACCURACY ON FIND FALLEN.

Method	Position Error (m) ↓	Type Acc. ↑
Modular Predictor [34]	2.41	0.32
Gradient Inversion	3.19	0.11
Ours Predictor	1.09	0.84

TABLE III
COMPARISON OF ANGLE, HEIGHT, AND DISTANCE CATEGORY
PREDICTION ACCURACY ON REALIMPACT DATASET.

Method	Angle Acc. ↑	Height Acc. ↑	Distance Acc. ↑
Chance	0.100	0.067	0.250
U-net+STFT [42]	0.825	0.902	0.972
CNN+waveform[43]	0.671	0.755	0.802
Resnet+STFT [40]	0.758	0.881	0.983
Ours	0.900	0.960	0.994

physically different rooms (32 study rooms and 32 kitchens). We evaluated the models’ performance on the test split identified by [34]. The audio was only available at the beginning of the test, and the agent would receive an RGB-D image at every subsequent step. For the available actions, we set `move forward` to 0.25 m and `rotate` to 30 degrees. The task was defined as follows: an embodied agent with an egocentric-view camera and microphone hears an unknown object fall somewhere in the room it is in (a study room or kitchen) as shown in Figure 3; the agent is then required to integrate audio-visual information to find which object has fallen and where it is, as efficiently as possible. Audio is recorded at 44.1 kHz in a two-channel configuration. We generate the STFT representation using a window and FFT

size of 512, a hop length of 128, and a Hann window. The PSD representation is generated using Welch’s method with a Hann window of size 256, an overlap of 128, and FFT size of 256.

We evaluate agents using three metrics: Success Rate, Success weighted by Path Length (SPL) [44], and Success weighted by Number of Actions (SNA) [33]. The Success Rate is calculated as the ratio of successful navigation trials to the total number of trials. A trial is considered successful if the agent explicitly executes action `found` when the distance between the agent and the object is less than 2 meters, the target physical object is visible in the agent’s view, and the number of actions executed so far is less than the maximum number of allowed steps (set to 200 in all tests). SPL is a metric that jointly considers the success rate and the path length to reach the goal from the starting point. SNA takes into account the number of actions and the success rate, penalizing collisions, rotations, and height adjustments taken to find the targets.

Result analysis. We evaluate the effectiveness of our proposed method against strong baselines as proposed in [34]. In Table IV, we find that our disentangled model based method can significantly outperform the previous modular planning baseline by 14% absolute in success rate. We further observe a significant decrease in the average length of the path and the number of actions taken by the agents. In addition, we test the method that firstly navigates to the location predicted by the encoder, and then utilizes the loss map for further exploration if the first attempt fails. In this case, we observe a comparable SR, but lower SPL and SNA metrics, which highlights the effectiveness of the loss map in exploration. We visualize the loss map and trajectories taken by the agent

TABLE IV

COMPARISON AGAINST BASELINE METHODS ON THE LOCALIZATION OF FALLEN OBJECTS. BASELINE RESULTS ARE TAKEN FROM [34].

Method	SR \uparrow	SPL \uparrow	SNA \uparrow
Decision Transformer [45]	0.17	0.12	0.14
PPO (Oracle found) [46]	0.19	0.15	0.14
SAVi [29]	0.23	0.16	0.10
Object-Goal [47]	0.22	0.18	0.17
Modular Planning [34]	0.41	0.27	0.25
Modular Planning + Loss Map	0.43	0.30	0.29
Modular Planning + Our Position	0.44	0.29	0.28
Modular Planning + Our Type	0.51	0.34	0.34
Ours (Full model)	0.57	0.38	0.37

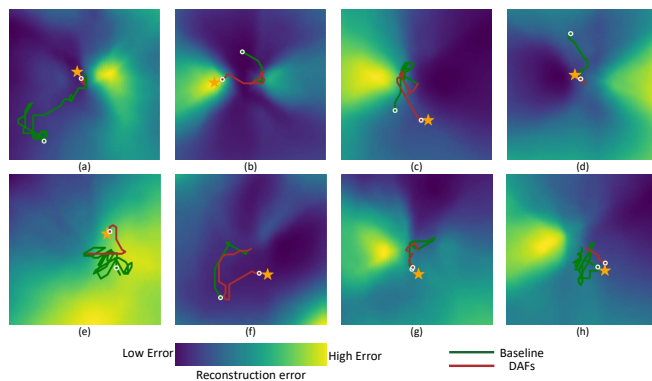


Fig. 4. **Comparison of agent trajectories.** We compare the agent trajectories using our method (Red) against the trajectories produced by the modular planning baseline (Green). The loss map uses dark blue to indicate regions of low error, while yellow is used to indicate regions of high error. This figure compares the uncertainty maps of various cases. Darker colors indicate lower values of position loss. The star (Gold) symbolizes the ground truth position of the fallen object. The end of each trajectory is circled in white for clarity. In (a)–(f), the baseline method fails to find the target, while our method succeeds. In (g)–(h), both methods find the target, but our method takes a shorter path.

in Figure 4. We observe in Figure 4 (a)–(f) that the modular planner often fails to find the target and attempts to find the object via random search, while our method can plan a direct route to the object. In Figure 4 (g)–(h), both methods find the target, but our method takes a shorter path. These results illustrate the superiority of our proposed approach over the baseline modular planner.

Ablation studies. The modular nature of our proposed method facilitates ablation studies that yield insight into the contribution of each individual component. We report results for ablations in Table IV. Beginning with the modular planning baseline, we find that augmenting the planner with a loss-map-guided object priority ranker yields a 2% increase in SR, a 3% increase in SPL, and 4% increase in SNA. This shows that the introduction of the uncertainty map can effectively improve the efficiency of agents searching for potential objects, reducing both the length of the path and the number of actions taken. Additionally, we replaced the sound

location and sound type predictors in modular planning with our predictor jointly trained with a generator. The improvement in the object type prediction accuracy was found to contribute more to the overall SR than the improvement in the position accuracy. This result corroborates the conclusion in [34] that accurately predicting object types from audio alone is a major challenge.

TABLE V

EVALUATION OF CROSS-SCENE PREDICTION FOR DAFs. COMPARED TO TABLE II, THERE IS A SMALL DECREASE IN ACCURACY.

Scene	Position Error (m) \downarrow	Type Acc. \uparrow
Kitchen to Study Room	1.17	0.81
Study Room to Kitchen	1.23	0.80

TABLE VI

EVALUATION OF CROSS-SCENE GENERALIZATION OF DIFFERENT METHODS. BASELINES ARE TAKEN FROM [34].

Method	Kitchen to Study Room			Study Room to Kitchen		
	SR \uparrow	SPL \uparrow	SNA \uparrow	SR \uparrow	SPL \uparrow	SNA \uparrow
PPO (Oracle found)	0.11	0.10	0.10	0.05	0.04	0.05
SAVi	0.20	0.11	0.09	0.19	0.14	0.11
Decision Transformer	0.07	0.06	0.06	0.08	0.06	0.07
Object Navigation	0.18	0.14	0.13	0.15	0.14	0.13
Modular Planning	0.34	0.23	0.20	0.35	0.22	0.19
Ours	0.52	0.38	0.37	0.48	0.32	0.32

Cross-Scene Generalization. To explicitly assess the generalization ability of our proposed method, we train and test on entirely different classes of rooms. In the first split, models are trained in the kitchens and tested in the study rooms. For the second split, models are trained in the study rooms and tested in the kitchens. The object property prediction results are reported in Table V. In both splits, the accuracy of positioning and predicting the type of object slightly decreased compared to that of the full-trained model. The planning results are reported in Table VI, where all models experience a degree of cross-scene performance drop. The success rate of the modular planning approach decreases by 7% in SR on the first split, while our method only decreases by 4%. Our proposed method still performs the best in both splits. This highlights that our method can not only generalize across room instances of the same type, but can also effectively generalize across rooms of a different type.

V. CONCLUSION

This paper presents an egocentric disentangled acoustic field framework that can generalize and reason across scenes. Joint inference of sound properties is implemented by using an explicit decomposition in the latent space. Furthermore, our approach is capable of generating multimodal uncertainty maps. Experiments on the TDW simulation platform demonstrate our disentangled acoustic field can improve the success rate for the localization of fallen objects. Consequently, our proposed method is a promising solution for sound localization and understanding in complex scenes.

REFERENCES

- [1] A. Luo, Y. Du, M. J. Tarr, J. B. Tenenbaum, A. Torralba, and C. Gan, "Learning neural acoustic fields," *arXiv preprint arXiv:2204.00628*, 2022.
- [2] A. Luo, T. Li, W.-H. Zhang, and T. S. Lee, "Surfgen: Adversarial 3d shape synthesis with explicit surface discriminators," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 16 238–16 248.
- [3] J. J. Park, P. Florence, J. Straub, R. Newcombe, and S. Lovegrove, "DeepSDF: Learning continuous signed distance functions for shape representation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 165–174.
- [4] D. Qu, C. Yan, D. Wang, J. Yin, Q. Chen, D. Xu, Y. Zhang, B. Zhao, and X. Li, "Implicit event-rgbd neural slam," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 19 584–19 594.
- [5] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis," *Communications of the ACM*, vol. 65, no. 1, pp. 99–106, 2021.
- [6] V. Sitzmann, M. Zollhöfer, and G. Wetzstein, "Scene representation networks: Continuous 3d-structure-aware neural scene representations," *Advances in Neural Information Processing Systems*, vol. 32, 2019.
- [7] Y. Xie, T. Takikawa, S. Saito, O. Litany, S. Yan, N. Khan, F. Tombari, J. Tompkin, V. Sitzmann, and S. Sridhar, "Neural fields in visual computing and beyond," in *Computer Graphics Forum*, vol. 41, no. 2. Wiley Online Library, 2022, pp. 641–676.
- [8] Y. Du, K. Collins, J. Tenenbaum, and V. Sitzmann, "Learning signal-agnostic manifolds of neural fields," *Advances in Neural Information Processing Systems*, vol. 34, pp. 8320–8331, 2021.
- [9] J. Yin, A. Li, T. Li, W. Yu, and D. Zou, "M2dgr: A multi-sensor and multi-scenario slam dataset for ground robots," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 2266–2273, 2021.
- [10] J. Yin, A. Li, W. Xi, W. Yu, and D. Zou, "Ground-fusion: A low-cost ground slam system robust to corner cases," *arXiv preprint arXiv:2402.14308*, 2024.
- [11] Z. Zhang, J. Wu, Q. Li, Z. Huang, J. B. Tenenbaum, and W. T. Freeman, "Inverting audio-visual simulation for shape and material perception," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 2536–2538.
- [12] A. Owens, J. Wu, J. H. McDermott, W. T. Freeman, and A. Torralba, "Ambient sound provides supervision for visual learning," in *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2016, pp. 801–816.
- [13] —, "Learning sight from sound: Ambient sound provides supervision for visual learning," *International Journal of Computer Vision*, vol. 126, pp. 1120–1137, 2018.
- [14] H. Zhu, M.-D. Luo, R. Wang, A.-H. Zheng, and R. He, "Deep audio-visual learning: A survey," *International Journal of Automation and Computing*, vol. 18, pp. 351–376, 2021.
- [15] A. Ephrat, I. Mosseri, O. Lang, T. Dekel, K. Wilson, A. Hassidim, W. T. Freeman, and M. Rubinstein, "Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation," in *Proceedings of ACM SIGGRAPH*, 2018.
- [16] H. Zhao, C. Gan, A. Rouditchenko, C. Vondrick, J. McDermott, and A. Torralba, "The sound of pixels," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 570–586.
- [17] R. Gao, R. Feris, and K. Grauman, "Learning to separate object sounds by watching unlabeled video," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 35–53.
- [18] R. Gao, C. Chen, Z. Al-Halah, C. Schissler, and K. Grauman, "Visualechoes: Spatial image representation learning through echolocation," in *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2020, pp. 658–676.
- [19] Z. Chen, X. Hu, and A. Owens, "Structure from silence: Learning scene structure from ambient sound," *arXiv preprint arXiv:2111.05846*, 2021.
- [20] R. Arandjelovic and A. Zisserman, "Look, listen and learn," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2017, pp. 609–617.
- [21] A. Owens and A. A. Efros, "Audio-visual scene analysis with self-supervised multisensory features," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 631–648.
- [22] C. Gan, D. Huang, P. Chen, J. B. Tenenbaum, and A. Torralba, "Foley music: Learning to generate music from videos," in *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2020, pp. 758–775.
- [23] A. Owens, P. Isola, J. McDermott, A. Torralba, E. H. Adelson, and W. T. Freeman, "Visually indicated sounds," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 2405–2413.
- [24] K. Su, X. Liu, and E. Shlizerman, "Audeo: Audio generation for a silent performance video," *Advances in Neural Information Processing Systems*, vol. 33, pp. 3325–3337, 2020.
- [25] L. Chen, S. Srivastava, Z. Duan, and C. Xu, "Deep cross-modal audio-visual generation," in *Proceedings of the Thematic Workshops of ACM Multimedia*, 2017, pp. 349–357.
- [26] T. Afouras, A. Owens, J. S. Chung, and A. Zisserman, "Self-supervised learning of audio-visual objects from video," in *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2020, pp. 208–224.
- [27] C. Gan, H. Zhao, P. Chen, D. Cox, and A. Torralba, "Self-supervised moving vehicle tracking with stereo sound," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 7053–7062.
- [28] R. Arandjelovic and A. Zisserman, "Objects that sound," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 435–451.
- [29] C. Chen, Z. Al-Halah, and K. Grauman, "Semantic audio-visual navigation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 15 516–15 525.
- [30] C. Chen, U. Jain, C. Schissler, S. V. A. Gari, Z. Al-Halah, V. K. Ithapu, P. Robinson, and K. Grauman, "Audio-visual embodied navigation," *environment*, vol. 97, p. 103, 2019.
- [31] C. Gan, Y. Zhang, J. Wu, B. Gong, and J. B. Tenenbaum, "Look, listen, and act: Towards audio-visual embodied navigation," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 9701–9707.
- [32] C. Chen, U. Jain, C. Schissler, S. V. A. Gari, Z. Al-Halah, V. K. Ithapu, P. Robinson, and K. Grauman, "Soundspaces: Audio-visual navigation in 3d environments," in *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2020, pp. 17–36.
- [33] C. Chen, S. Majumder, Z. Al-Halah, R. Gao, S. K. Ramakrishnan, and K. Grauman, "Learning to set waypoints for audio-visual navigation," *arXiv preprint arXiv:2008.09622*, 2020.
- [34] C. Gan, Y. Gu, S. Zhou, J. Schwartz, S. Alter, J. Traer, D. Gutfreund, J. B. Tenenbaum, J. H. McDermott, and A. Torralba, "Finding fallen objects via asynchronous audio-visual integration," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 10 523–10 533.
- [35] C. Gan, J. Schwartz, S. Alter, D. Mrowca, M. Schrimpf, J. Traer, J. De Freitas, J. Kubilius, A. Bhandwadar, N. Haber, *et al.*, "Three-world: A platform for interactive multi-modal physical simulation," *arXiv preprint arXiv:2007.04954*, 2020.
- [36] Google, "Google resonance audio," <https://resonance-audio.github.io/resonance-audio/develop/overview.html>, 2018.
- [37] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.
- [38] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2017, pp. 2961–2969.
- [39] P. E. Hart, N. J. Nilsson, and B. Raphael, "A formal basis for the heuristic determination of minimum cost paths," *IEEE transactions on Systems Man and Cybernetics*, vol. 4, no. 2, pp. 100–107, 1968.
- [40] R. Gao, Z. Si, Y.-Y. Chang, S. Clarke, J. Bohg, L. Fei-Fei, W. Yuan, and J. Wu, "Objectfolder 2.0: A multisensory object dataset for sim2real transfer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 10 598–10 608.
- [41] S. Clarke, R. Gao, M. Wang, M. Rau, J. Xu, J.-H. Wang, D. L. James, and J. Wu, "RealImpact: A dataset of impact sound fields for real objects," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 1516–1525.
- [42] S. Patel, M. Zawodniok, and J. Benesty, "Dcase 2020 task 3: A single stage fully convolutional neural network for sound source localization and detection," *DCASE2020 Challenge*, 2020.
- [43] Y. He, N. Trigoni, and A. Markham, "Sounddet: Polyphonic moving sound event detection and localization from raw waveform," in

International Conference on Machine Learning. PMLR, 2021, pp. 4160–4170.

- [44] P. Anderson, A. Chang, D. S. Chaplot, A. Dosovitskiy, S. Gupta, V. Koltun, J. Kosecka, J. Malik, R. Mottaghi, M. Savva, *et al.*, “On evaluation of embodied navigation agents,” *arXiv preprint arXiv:1807.06757*, 2018.
- [45] L. Chen, K. Lu, A. Rajeswaran, K. Lee, A. Grover, M. Laskin, P. Abbeel, A. Srinivas, and I. Mordatch, “Decision transformer: Reinforcement learning via sequence modeling,” *Advances in neural information processing systems*, vol. 34, pp. 15 084–15 097, 2021.
- [46] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, “Proximal policy optimization algorithms,” *arXiv preprint arXiv:1707.06347*, 2017.
- [47] D. S. Chaplot, D. P. Gandhi, A. Gupta, and R. R. Salakhutdinov, “Object goal navigation using goal-oriented semantic exploration,” *Advances in Neural Information Processing Systems*, vol. 33, pp. 4247–4258, 2020.

APPENDIX

A. Navigation Failure Case Analysis

Despite significant performance improvements upon baseline methods in the navigation task, our method still may sometimes fail to find the fallen object. Upon analyzing some failure cases, we discover that inaccurate semantic segmentation is one major problem. In such scenes, even if the target position and type are accurately predicted by our audio network, the agent would not be able to find the object. As illustrated in Figure 5, some segmentation failures are due to the object being too small, or its color being too close to the background color. Additionally, distractor objects may cause the agent to use the `found` command on the wrong object. We show two trajectories where our method cannot find the object or uses a long path. Figure 6 (a) shows a case where multiple distractors of the same kind as the fallen object are in the agent’s view, and thus both the baseline and our method fail to navigate to the target location. As seen in Figure 6 (b), the introduction of a loss map does not necessarily ensure that the agent takes the shortest path. Our method succeeds in finding the target, but it takes a longer path than the baseline.

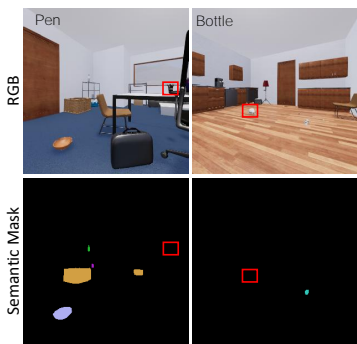


Fig. 5. **Failure from visual branch.** The visual branch is learned independently of the auditory branch. Semantic segmentation errors can occur when objects are visually small or of low contrast. Future work can explore the contrastive learning of joint audio-visual representations.

B. Choice of output representation

In the main paper, we choose to utilize power spectral density (PSD) as the choice of output representation, instead of the short-time Fourier transform (STFT). The PSD representation captures the power in each specific frequency,

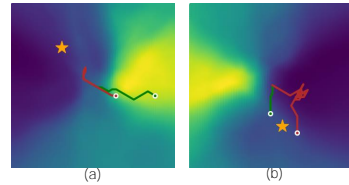


Fig. 6. **Trajectory of failure cases.** (a) Both our method and the modular planning baseline execute the `found` command on the wrong object. (b) Our method takes a longer path than the baseline method and searches a low error region of the loss map.

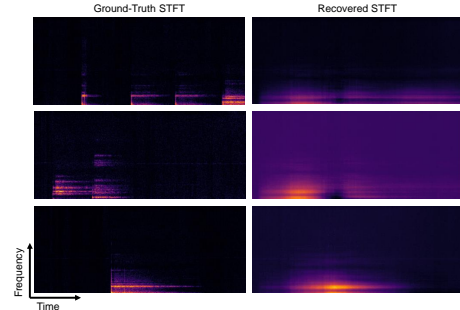


Fig. 7. **Visualization of STFT Reconstruction.** **Left:** The ground-truth log-STFT of an impact sound. **Right:** The log-STFT recovered by our network. In the three examples shown here, we find that the network struggles to model the irregular temporal caps present in the STFT.

but unlike STFT, it discards the temporal information (PSD is similar to an average pooling across time of the squared STFT magnitude).

In this experiment, we seek to reconstruct the STFT. We utilize the network parameterization proposed by Neural Acoustic Fields (NAFs). The STFT was supervised with an MSE loss. After training, we evaluate the two models on a test set of 100 instances. The results of the object type and location predictions are shown in Table VII, where the predictor supervised by PSD reconstruction achieves higher accuracy in both object type and position. Figures 7 and 8 compare the ground-truth and recovered values of STFT and PSD, respectively. These results show that the low-dimensional PSD is easier to reconstruct with high quality than STFT.

TABLE VII
COMPARISON OF POSITION ERROR AND TYPE PREDICTION ACCURACY DEPENDING ON THE OUTPUT REPRESENTATION. MODEL-P DENOTES THE MODEL SUPERVISED BY PSD RECONSTRUCTION, WHILE MODEL-S DENOTES THE MODEL SUPERVISED BY STFT RECONSTRUCTION.

Model’s output representation	Position Error (m)↓	Type Acc.↑
STFT	2.03	0.57
PSD	1.44	0.73

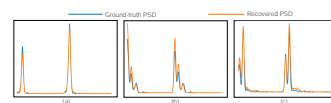


Fig. 8. **Visualization of PSD Reconstruction.** We visualize the ground truth PSD in blue, while the network predicted PSD is shown in orange. In the three examples (a)–(c), we find that the network can reconstruct the PSD with low error.