Beyond Visual Field of View: Perceiving 3D Environment with Echoes and Vision

Lingyu Zhu
Tampere University
lingyu.zhu@tuni.fi

Esa Rahtu
Tampere University
esa.rahtu@tuni.fi

Hang Zhao
Tsinghua University
zhaohang0124@gmail.com

Abstract

This paper focuses on perceiving and navigating 3D environments using echoes and RGB image. In particular, we perform depth estimation by fusing RGB image with echoes, received from multiple orientations. Unlike previous works, we go beyond the field of view of the RGB image and estimate dense depth maps for substantially larger parts of the environment. Moreover, we study to leverage echoes and visual observations for robot navigation. We show that the echoes provide holistic information about the 3D structures complementing the visual observations. We compare the proposed methods against recent baselines using two sets of realistic 3D environments: Replica and Matterport3D.

1. Introduction

The structure of a 3D environment is commonly inferred using RGB images or active depth sensors [4, 16]. While they provide detailed information, the observations are usually limited to a small field of view (FoV). This limitation can be compensated by installing multiple cameras or physically moving the device. However, such procedure requires processing multiple images, which might be unnecessarily heavy for completing the required tasks (e.g., navigation).

Recent works [8, 11, 14, 17, 21] have shown that jointly utilizing the audio-visual observations can substantially enhance the spatial reasoning of physical space. However, these methods mostly focus on enhancing the prediction in the same area as the RGB covers. One advantage of echoes is that echoes naturally have a wider "field of view" than the RGB observation. Instead of focusing on inside the RGB FoV, one can benefit most from outside the RGB FoV when leveraging echoes. These motivate us to study utilizing echo information to go beyond the RGB FoV and to extend the prediction for wider FoV.

In summary, our key contributions include i) an end-to-end neural network that learns to take advantage of echoes received from multiple orientations and RGB image for better depth estimation; ii) leveraging echoes to extend depth prediction over RGB FoV; iii) introducing a novel Point-Goal echo navigation that directly utilizes binaural echoes to holistically perceive the physical space. It outperforms the method of using RGB; iv) Fusing the echoes to visual observations further improves the navigation performance. Without adding more cameras and additional processing, utilizing echoes helps to overcome the limitations of narrow visual FoV and to obtain better understanding of the 3D environment.

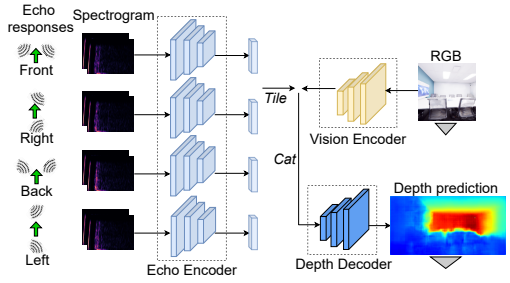

Figure 1. Framework of depth estimation using echoes and RGB.

2. Predicting Wide FoV Depth Maps from Echoes and RGB

In this section, we propose to estimate a large FoV depth map using echoes and a narrow FoV RGB.

2.1. Depth Estimation From Echoes

In order to take advantage of echoes from multiple orientations, we propose a framework in Fig. 1 (omitting the vision encoder). It contains four echo encoders and a depth predictor. The echo encoders share parameters. Each echo encoder is a convolutional neural network. It is composed of three continuous blocks of $\{Conv, BatchNorm, ReLU\}$. A $Flatten$ and $Conv$ layer with kernel 1×1 are appended at the end to convert the echo features into a vector of size 512×1. The input of each echo encoder is represented as two-channel Frequency-Time (F-T) spectrograms that are obtained by applying Short-time Fourier Transform (STFT) to the received binaural echo responses. The depth decoder consists

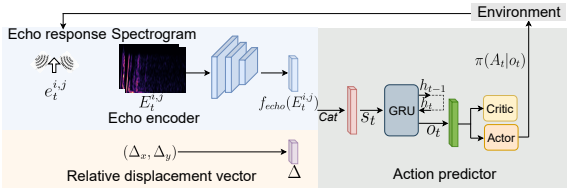

Figure 2. The architecture of the PointGoal echo navigation.

of 6 blocks of $\{ConvTranspose, BatchNorm, ReLU\}$ and a following $\{ConvTranspose, Sigmoid\}$ for projecting features into one channel depth prediction within range [0, 1].

2.2. Depth Estimation from Echoes and RGB

Estimating depth maps beyond the visual FoV: The architecture in Fig. 1 fuses the echoes into RGB for predicting a wide FoV depth. The vision encoder consists of five $Conv$ layers. Each layer is followed by a $BatchNorm$ and $ReLU$. We tile the encoded echo feature vector to match the spatial dimension of visual features and then concatenate the echo and visual feature maps along the channel dimension to pass to the depth decoder.

Given an RGB image with FoV θ, e.g., 120°, we augment an input RGB image of a smaller FoV $\theta^{'}$ by masking an "unseen" region from two sides of this full RGB image as zeros. The new $width^{'}$ corresponding to the FoV $\theta^{'}$ is computed through,

$$width^{'} = width * tan(\frac{\theta^{'} * \pi}{360^{\circ}})/tan(\frac{\theta * \pi}{360^{\circ}}), \qquad (1)$$

where the $width$ represents the original RGB width corresponding to an entire FoV θ. The computed $width^{'}$ corresponds to a specific smaller FoV $\theta^{'} \in (0, \theta]$. Note that we only consider the FoV changes in horizon.

Extending depth prediction to completely unseen areas: Predicting to extend the RGB neighborhood FoV potentially benefit from the visual similarity and extension of environmental surfaces. One might be interested in using the information from completely unseen areas to predict a depth map, for instance, predicting the "front" depth using the RGB from "left", "right", or "back" side. Thus, we study a more extreme case, that is, to predict the depth at the forward orientation when giving echoes and RGB image from a total side or opposite direction. Note that there is no overlap among the RGB observations (FoV 90°).

3. Navigating Using Echoes and RGB

We introduce PointGoal echo navigation (Fig. 2) to directly use echoes to perceive the spatial cues of physical space for 3D navigation. Moreover, we take advantage of audio-visual learning by fusing echoes to visual observations for better embodied 3D navigation.

Table 1. Comparing depth estimation performance inside RGB FoV between our proposed models and baseline methods using Replica [20] and Matterport3D [3] datasets.

Dataset	Method	rmse	rel	log10	$\delta_{1.25}$	$\delta_{1.25^2}$	$\delta_{1.25^3}$
Replica	Average [11]	1.070	0.791	0.230	0.235	0.509	0.750
	Echo2Depth [11]	0.969	0.753	0.204	0.441	0.631	0.752
	RGB2Depth [11]	0.374	0.202	0.076	0.749	0.883	0.945
	VisualEchoes [11]	0.346	0.172	0.068	0.798	0.905	0.950
	Materials [17]	0.249	0.118	0.046	0.869	0.943	0.970
	Our(Echoes)	0.797	0.534	0.171	0.544	0.708	0.802
	Our(Echoes+RGB)	0.294	0.166	0.060	0.814	0.912	0.958
MP3D	Average [11]	1.913	0.714	0.237	0.264	0.538	0.697
	Echo2Depth [11]	1.778	0.507	0.192	0.464	0.642	0.759
	RGB2Depth [11]	1.090	0.260	0.111	0.592	0.802	0.910
	VisualEchoes [11]	0.998	0.193	0.083	0.711	0.878	0.945
	Materials [17]	0.950	0.175	0.079	0.733	0.886	0.948
	Our(Echoes)	1.535	0.465	0.184	0.476	0.664	0.781
	Our(Echoes+RGB)	0.777	0.161	0.069	0.775	0.874	0.943

3.1. PointGoal Echo Navigation Task Setup

We introduce PointGoal echo navigation that employs binaural echoes for embodied 3D navigation. Given a point goal defined by a displacement vector (Δ_x, Δ_y) relative to the starting position of the agent, the task of PointGoal echo navigation is to let the agent navigate to the point goal by keeping receiving binaural echoes while moving. Note that there is no map of the scene is available to the agent. The agent needs to avoid obstacles, navigate, and reach the target goal by perceiving spatial cues using sensory input. The sensory inputs are GPS and binaural echoes. An idealized GPS sensor [4, 12, 15, 18] offers the relative location of the target goal. To emit the omnidirectional sweep signal and receive the binaural echoes, we emulate one speaker and a microphone array (four pairs microphones) on the agent. The navigation actions consist of four actuations: *MoveForward*, *TurnLeft*, *TurnRight*, and *Stop*.

3.2. Echo Navigation Network

We introduce echoes as the agent input and study a policy for mapping the sensory input to agent actions by adopting deep reinforcement learning with Proximal Policy Optimization (PPO) [19].

In Fig. 2, the echo encoder is a convolutional neural network that stacks three $Conv$ layers following $ReLU$ operations. We apply STFT on top of the received binaural echo responses $e_t^{i,j}$ to compute the binaural echo spectrograms $E_t^{i,j}$ as the input of the echo encoder f_{echo}. The $e_t^{i,j}$ and $E_t^{i,j}$ indicate, at time step t, the binaural echo response and its spectrograms of the agent at location i with orientation j. The following $Flatten$ and $Fully\ connected$ layers map the echo features into a feature vector of size 512×1, which

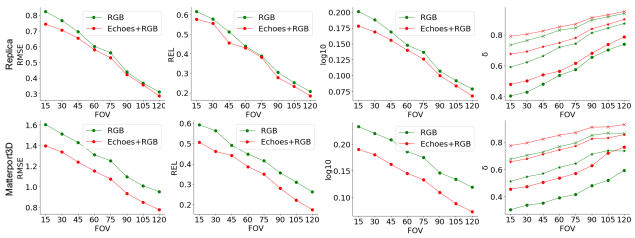

Figure 3. Depth prediction using RGB FoV $\in (0, 120]$, w/o echoes (green) and w/ echoes (red). For δ (last column), the curves with $\bullet$, $*$, and x denote the $\delta_{1.25}$, $\delta_{1.25^2}$, and $\delta_{1.25^3}$, respectively.

preserves the room geometry and agent's position. We also study to better navigate with audio-visual learning by leveraging echoes to visual observations. A neural network with three *Conv* layers and a *Fully connected* layer (each layer follows with a *ReLU* operation) is applied to process the visual observation (not visible in Fig. 2).

The echo features $f_{echo}(E_t^{i,j})$ carry important spatial information and make the agent aware of its position and the room geometry, which is beneficial for the agent to execute an action towards the point goal. The Gated Recurrent Unit (GRU) [7, 9] module takes as input the concatenation of the echo feature vector $f_{echo}(E_t^{i,j})$ and the given GPS displacement vector Δ to recursively process each symbol while maintaining its internal hidden state h. The following is a reinforcement learning policy equipped with an actor-critic architecture. It produces a probability distribution $\pi(A_t|o_t)$ over possible actions by operating on the predicted agent state o_t from the GRU module,

$$s_t = Cat(f_{echo}(E_t^{i,j}), \Delta), \tag{2}$$

$$o_t, h_t = f_{gru}(s_t, h_{t-1}) \tag{3}$$

where the *Cat*, f_{echo}, and f_{gru} denote concatenation, echo encoder, and GRU operation, respectively. A_t represents the candidate actions from the action space. We sample an action a_t from A_t according to the policy's predicted probability distribution.

4. Experiments

We evaluate the proposed methods using depth estimation and robot navigation. We report the depth prediction with standard metrics of root mean squared error (RMSE), mean relative error (REL), mean log10 error (log10), and the thresholded accuracy of $\{\delta_{1.25}, \delta_{1.25^2}, \delta_{1.25^3}\}$ [10, 13]. For navigation, we evaluate with the success rate normalized by inverse path length (SPL) [1].

4.1. Datasets

SoundSpaces [5] is a realistic acoustic simulation platform that augments the Habitat simulator [18]. Habitat is an open-source 3D simulator that supports fast rendering for multiple datasets on RGB, depth, and semantics. SoundSpaces enables audio rendering based on geometrical acoustic simulations for two sets of publicly available 3D environments Replica [20] and Matterport3D [3]. Replica is a dataset with 3D meshes from real-world scans of 18 scenes ranging in area from 9.5 to 141.5 m^2. The Matterport3D dataset consists of 85 large environments ranging from 53.1 to 2921.3 m^2, which are real-world indoor environments with 3D meshes and image scans.

4.2. Depth Estimation from Echoes and RGB

Comparison with state-of-the-art: We start by comparing our models against competitive baselines of Average [11], Echo2Depth [11], RGB2Depth [11], VisualEchoes [11], and Materials [17] to estimate the depth inside RGB FoV (90°) in Table 1. We adopt the same experimental setup as [11, 17]. We find that combining echoes received from multiple orientations achieves better results than using one pair alone. Our method that combines echoes received from four orientations performs better than Echo2Depth [11] with a large margin. With the presence of target orientation RGB image, our proposed approach achieves an improvement of 15.0% (Replica) and 22.1% (Matterport3D) over VisualEchoes [11]. Furthermore, we observe that the method Materials [17] performs the best on Replica dataset while only attaining similar results as VisualEchoes [11] on Matterport3D. This may explain that the material cues brought from the pretrained material approach [2] have a dominant impact on depth prediction on Replica. However, for large Matterport3D environment scenes, its influence declines. Remarkably, our model achieves state-of-the-art results on Matterport3D dataset, overwhelming Materials [17] around 18.2% on RMSE. It is worth noting that the Materials [17] model has 316.9M parameters in comparison to our 21.7M. These indicate that our proposed methods better perceive the geometrical information of 3D environments.

Combining echoes with RGB to estimate a wide FoV depth: We examine our model in Fig. 1 (w/o and w/ echoes) for depth extension (FoV 120°) using echoes and RGB (FoV $\in \{15°, 30°, 45°, 60°, 75°, 90°, 105°, 120°\}$). We observe from Fig. 3 that associating echoes outperforms the counterpart results (w/o echoes) over all the FoVs. Especially for Replica, the improvement gets smaller when increasing the RGB FoV. Thus, expanding RGB FoV to apply to echoes does not bring large performance gain. This shows that the echo serves as a strong spatial cue when it goes to the region where RGB is unavailable. Interestingly, when enlarging the RGB FoV for Matterport3D, the model using echoes receives a relatively stable improvement over the RGB-based model. It is likely because the Matterport3D contains large 3D environment scenes, and the RGB cap-

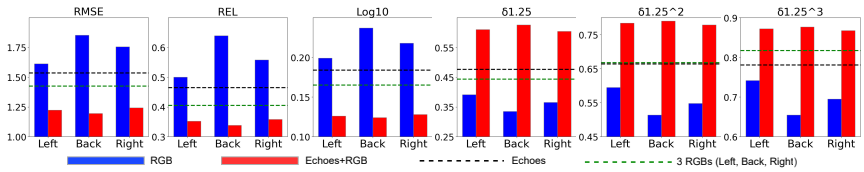

Figure 4. Bar charts of predicting "front" depth by leveraging echoes with "left", "back", or "right" side RGB image using Matterport3D.

tures important geometric structures for large scenes.

Leveraging echoes and RGB to predict depth of completely unseen areas: It is a more challenging problem when there is no overlap between the input RGB and target depth. Fig. 4 visualizes the depth prediction metrics for using echoes and the RGB images from sideways ("left" and "right") and "back". The input RGB and predicted depth are of FoV 90°. The similarity and extension of the visual surfaces between the forward and backward RGB images is comparatively very low. This is disclosed by the worse performance of the "back" blue bars compared to the "left" and "right". The model using echoes alone (dashed black line) performs better than using the RGB image (blue bars). We also experiment with the depth estimation of target orientation using RGB images from the three rest orientations. For instance, we use the RGB images from the "left", "right", and "back" sides to predict the front depth. Its result is shown as the dashed green line in Fig. 4, which indicates that investing in additional cameras can bring performance gain but increase substantial computing complexity.

Fusing echoes into the RGB (red bars) attains superior improvements. For all metrics, the red bars surpass the rest methods by a large margin. These reflect the efficacy of fusing echoes into RGB for exploiting the geometrical information. Specifically, we observe that, after fusing echoes into RGB, the performance differences among "left", "right", and "back" get smaller for all the metrics. This is interesting because it suggests that the echoes capture complementary spatial information for each orientation, showing echoes contain robust geometrical cues when going outside the RGB FoV.

4.3. Navigating Using Echoes and RGB

We consider three visual sensing learning based baselines which predict action using GPS sensor together with the visual input of no visual observation, raw RGB image, and depth image. Similar to [4, 6, 12, 15, 18], agents are allowed a time horizon of 500 actions for all tasks. Table 2 summarizes the navigation performance of SPL in comparison with baselines using the test environments from Replica and Matterport3D. We observe that adding RGB sensor improves the results over the blind (only GPS) option.

Compared to the results of using visual sensory input, the performance of directly utilizing echoes (Replica: 0.547 and Matterport3D: 0.474) stands between the results of applying RGB and depth. It reveals that the PointGoal echo navigation captures more vital spatial cues than the PointGoal RGB navigation. These findings may result from i) the fact that echoes naturally capture a holistic understanding of the 3D environment, thus have great spatial perception of the physical space; and ii) perceiving geometrical information of 3D environment to predict an agent action towards a point goal is technically a coarse-grained classification task for which echoes are a strong cue already.

Table 2. Navigation performance in SPL using echoes and vision.

Echoes	RGB-90°	Depth-90°	Depth-120°	Replica	MP3D
✗	✗	✗	✗	0.491	0.425
✗	✓	✗	✗	0.526	0.448
✗	✗	✓	✗	0.599	0.531
✓	✗	✗	✗	0.547	0.474
✓	✓	✗	✗	0.563	0.490
✓	✗	✓	✗	0.613	0.553
✗	✗	✗	✓	0.624	0.546
✓	✗	✗	✓	0.627	0.562

In Table 2, combing echoes with the visual sensory input further improves the result, which indicates the efficacy of audio-visual learning. The strong geometric structure contained in echoes and depth makes the method of *Echoes+Depth-90°* outperform *Echoes* and *Echoes+RGB-90°*. In order to verify whether the models benefit from the holistic understanding of the environment by echoes, we enlarge the FoV of input depth to fuse with echoes for navigation. When increasing the depth FoV from 90° to 120°, the improvement gain of *Echoes+Depth-120°* over *Depth-120°* is smaller than the experiments using FoV 90°. This reflects our observation from depth estimation in Fig. 3, 4 and suggests echoes can perceive strong geometrical cues when going outside the visual FoV.

5. Conclusion

Our work sheds light on utilizing echoes to extend geometrical understanding of physical space over visual observations. We leverage echoes to RGB for predicting depth of a substantially large FoV. Besides, we introduce PointGoal echo navigation, which outperforms PointGoal RGB navigation. Leveraging echoes to visual observation further improves performance.

References

[1] Peter Anderson, Angel Chang, Devendra Singh Chaplot, Alexey Dosovitskiy, Saurabh Gupta, Vladlen Koltun, Jana Kosecka, Jitendra Malik, Roozbeh Mottaghi, Manolis Savva, et al. On evaluation of embodied navigation agents. *arXiv preprint arXiv:1807.06757*, 2018. 3

[2] Sean Bell, Paul Upchurch, Noah Snavely, and Kavita Bala. Material recognition in the wild with the materials in context database. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3479–3487, 2015. 3

[3] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. *arXiv preprint arXiv:1709.06158*, 2017. 2, 3

[4] Devendra Singh Chaplot, Dhiraj Gandhi, Saurabh Gupta, Abhinav Gupta, and Ruslan Salakhutdinov. Learning to explore using active neural slam. In *International Conference on Learning Representations (ICLR)*, 2020. 1, 2, 4

[5] 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 *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*, pages 17–36. Springer, 2020. 3

[6] Tao Chen, Saurabh Gupta, and Abhinav Gupta. Learning exploration policies for navigation. In *International Conference on Learning Representations*, 2019. 4

[7] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014. 3

[8] Jesper Haahr Christensen, Sascha Hornauer, and X Yu Stella. Batvision: Learning to see 3d spatial layout with two ears. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1581–1587. IEEE, 2020. 1

[9] Junyoung Chung, Kyle Kastner, Laurent Dinh, Kratarth Goel, Aaron C Courville, and Yoshua Bengio. A recurrent latent variable model for sequential data. *Advances in neural information processing systems*, 28:2980–2988, 2015. 3

[10] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. *arXiv preprint arXiv:1406.2283*, 2014. 3

[11] Ruohan Gao, Changan Chen, Ziad Al-Halah, Carl Schissler, and Kristen Grauman. Visualechoes: Spatial image representation learning through echolocation. In *European Conference on Computer Vision*, pages 658–676. Springer, 2020. 1, 2, 3

[12] Daniel Gordon, Abhishek Kadian, Devi Parikh, Judy Hoffman, and Dhruv Batra. Splitnet: Sim2sim and task2task transfer for embodied visual navigation. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019. 2, 4

[13] Junjie Hu, Mete Ozay, Yan Zhang, and Takayuki Okatani. Revisiting single image depth estimation: Toward higher resolution maps with accurate object boundaries. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1043–1051. IEEE, 2019. 3

[14] Hansung Kim, Luca Remaggi, Philip JB Jackson, Filippo Maria Fazi, and Adrian Hilton. 3d room geometry reconstruction using audio-visual sensors. In *2017 International Conference on 3D Vision (3DV)*, pages 621–629. IEEE, 2017. 1

[15] Noriyuki Kojima and Jia Deng. To learn or not to learn: Analyzing the role of learning for navigation in virtual environments. *arXiv preprint arXiv:1907.11770*, 2019. 2, 4

[16] Fayao Liu, Chunhua Shen, and Guosheng Lin. Deep convolutional neural fields for depth estimation from a single image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5162–5170, 2015. 1

[17] Kranti Kumar Parida, Siddharth Srivastava, and Gaurav Sharma. Beyond image to depth: Improving depth prediction using echoes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8268–8277, 2021. 1, 2, 3

[18] Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, et al. Habitat: A platform for embodied ai research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9339–9347, 2019. 2, 3, 4

[19] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. 2

[20] Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, et al. The replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*, 2019. 2, 3

[21] Mao Ye, Yu Zhang, Ruigang Yang, and Dinesh Manocha. 3d reconstruction in the presence of glasses by acoustic and stereo fusion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4885–4893, 2015. 1