Supplementary Material for
Beyond Image to Depth: Improving Depth Prediction using Echoes

Kranti Kumar Parida[1] Siddharth Srivastava[2] Gaurav Sharma[1,3]

[1] IIT Kanpur [2] CDAC Noida [3] TensorTour Inc.

{kranti, grv}@cse.iitk.ac.in, siddharthsrivastava@cdac.in

1. Dataset Details

We use two datasets Replica [6] and Matterport3D [1] for our experiments. Both the datasets are rendered using an open source 3D simulator, Habitat [5]. To obtain echoes on both the datasets, we use the simulations from Soundspaces [2]. Soundspaces augments the simulator by providing realistic audio simulations for the scenes by considering room geometry and materials in the room.

1.1. Simulating Echoes

We use the procedure outlined below to obtain echoes on both Replica and Matterport3D dataset. Soundspaces performs acoustic simulation in two steps as follows.

Step 1. The visual scene from the respective dataset is subdivided into grids. The grids are divided along navigable points so that an agent can be placed there. Then the Room Impulse Response (RIR) is computed between each pair of points using audio ray tracing [7]. Each pair denotes a combination of source and receiver which send the audio signal and receive the echoes respectively.

Step 2. The echoes are obtained by convolving the input audio signal with the RIR computed in the previous step.

Following Soundspaces, we use the RIR between each pair of point at four orientations ($0°$, $90°$, $180°$, $270°$). For the proposed method, we place the source and receiver at the same point and use the resulting RIR. In addition, following [3], we use the source audio signal as a 3 ms sweep signal spanning the human hearing range (20Hz to 20kHz). We obtain the echo response by convolving the corresponding source audio signal with the RIRs obtained previously. Further, the sampling rate for the source and received audio (echoes) are 44.1 kHz and 16 kHz for Replica and Matterport3D respectively.

1.2. Visual Scenes

We now provide details on the scenes used from each dataset along with the train and test details.

Replica dataset. We use all the 18 scenes from Replica having 6960 points in total, from 1740 images and 4 orientations. Following [3], we use a train set consisting of 5496 points and 15 scenes. The test set consists of 1464 points from 3 scenes. As a validation set is not defined for Replica, we use a small subset of points from train set for tuning the network parameters. Then the parameters are fixed, and the entire train set is used training the network.

Matterport3D dataset. Matterport3D consists of 90 scenes. Soundspaces provides RIR for 85 of these scenes. Further, we discard another 8 scenes which have none or a very few navigable points. This results in a dataset with 77 scenes which we use as our final dataset. These 77 scenes contain $67,376$ points from $16,844$ and 4 orientations. The dataset is then split into train, validation and test sets. The train set consists of $40,176$ points and 59 scenes. The validation set consists of $13,592$ points and 10 scenes. The test set consists of $13,602$ points and 8 scenes.

2. Network Architecture and Parameters

We now provide the detailed architecture of each subnetwork from the proposed method.

Echo Net. It is an encoder-decoder network. The encoder is inspired from [3] and consists of a convolutional neural network having 3 layers with filter dimensions 8×8, 4×4, 3×3 and stride 4×4, 2×2 and 1×1 respectively for each layer. The number of output filters in each layers are 32, 64 and 8 respectively. Finally, we use a 1×1 conv layer to convert arbitrary sized feature dimension into a 512 dimensional feature vector.

The decoder consists of 7 fractionally strided convolutional layers with filter size, stride and padding of 4,2 and 1 respectively. The number of output filters for the 7 layers are $512, 256, 128, 64, 32, 16$ and 1 respectively. We use BatchNorm and RELU non-linearity after each layer of the network.

Visual Net. It consists of an encoder decoder network. The encoder consists of a convolutional neural network with 5 layers. For each layer, the filter size is 4, the stride is 2 and padding is 1. The 5 layers have $64, 128, 256, 512, 512$

number of output filters respectively. We use LeakyRELU with negative slope of 0.2 and BatchNorm after each layer.

Similarly for the decoder we use 5 fractionally strided convolutional layers with output filters of size $512, 256, 128, 64, 1$ respectively. We also use skip connections and concat the features from the corresponding encoder layer with decoder output to get the final feature map from the decoder. We use BatchNorm and RELU after each layer.

Material Net. We use the first five convolution blocks of the ResNet-18 [4]. The first layer has a filter size of 7×7 and all subsequent layer have filters of size 3×3. The number of output filters at each layer are $64, 64, 128, 256, 512$ respectively.

Attention Net. We use five fractionally strided convolutional layers with output filter sizes of $512, 256, 128, 64, 1$ respectively for each layer. We use filter size, stride and padding to be $4, 2, 1$ respectively.

3. Implementation Details

Input. The input to the Visual Net and Material Net is a 128x128 RGB image. We also perform image augmentation by randomly jittering color, contrast and brightness of the image.

The input to the Echo Net is a spectrogram from the simulated echoes. For obtaining the spectrogram, we first convert the time domain audio signal into Short Time Fourier Transform using Hanning window with a fixed window length, hop length and frequency points. We use a two channel audio with duration of 60ms.

For Replica, we use an audio signal of 44.1kHz and convert it to a $2 \times 257 \times 166$ spectrogram using a window length of 64, hop length of 16 and 512 frequency points.

For Matterport3D, we use an audio signal of 16kHz and convert it to a $2 \times 257 \times 121$ spectrogram using a window length of 32, hop length of 8 and 512 frequency points.

Additional Parameters. We train the network on both the datasets using Adam optimizer with learning rate of $1e-4$, momentum of 0.9 and weight decay of $5e-4$. We use batch size of 128 for Replica and 64 for Matterport3D.

4. Evaluation Metrics

We use following metrics to evaluate our result.

We denote the predicted depth and ground truth depth as $\hat{\mathbf{D}}(p)$ and $\mathbf{D}(p)$ for every point p. We further use only those points that have valid depth value, i.e. the missing values and the points having zero depth value in $\mathbf{D}$ are ignored. We denote such valid points as $|\mathbf{D}|$.

- Root Mean Square Error:

$$\sqrt{\frac{1}{|\mathbf{D}|}\sum_{p\in\mathbf{D}} |\hat{\mathbf{D}}(p) - \mathbf{D}(p)\|^2} \tag{1}$$

- Mean absolute relative error:

$$\frac{1}{|\mathbf{D}|}\sum_{p\in\mathbf{D}} \frac{\hat{\mathbf{D}}(p) - \mathbf{D}(p)}{\hat{\mathbf{D}}(p)} \tag{2}$$

- Mean $\log_{10}$ error:

$$\frac{1}{|\mathbf{D}|}\sum_{p\in\mathbf{D}} \log_{10}(\hat{\mathbf{D}}(p)) - \log_{10}(\mathbf{D}(p)) \tag{3}$$

- δ_t is the percentage of pixels within the error range t. We define the error range as mentioned below

$$max(\frac{\hat{\mathbf{D}}(p)}{\mathbf{D}(p)}, \frac{\mathbf{D}(p)}{\hat{\mathbf{D}}(p)}) < t \tag{4}$$

where $t \in \{1.25, 1.25^2, 1.25^3\}$.

5. More Qualitative Results

We give more qualitative results of depth estimation using various techniques on Replica and Matterport3D datasets in Fig. 1 and Fig. 2 respectively. The visualizations of the attention maps from Echo Net and Visual Net are shown in Fig. 3 (Replica) and Fig.4 (Matterport3D).

References

[1] 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. *International Conference on 3D Vision (3DV)*, 2017. 1

[2] 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. *ECCV*, 2020. 1

[3] Ruohan Gao, Changan Chen, Ziad Al-Halah, Carl Schissler, and Kristen Grauman. Visualechoes: Spatial image representation learning through echolocation. *ECCV*, 2020. 1

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2

[5] 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 International Conference on Computer Vision*, pages 9339–9347, 2019. 1

[6] 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. 1

[7] Eric Veach and Leonidas Guibas. Bidirectional estimators for light transport. In *Photorealistic Rendering Techniques*, pages 145–167. Springer, 1995. 1

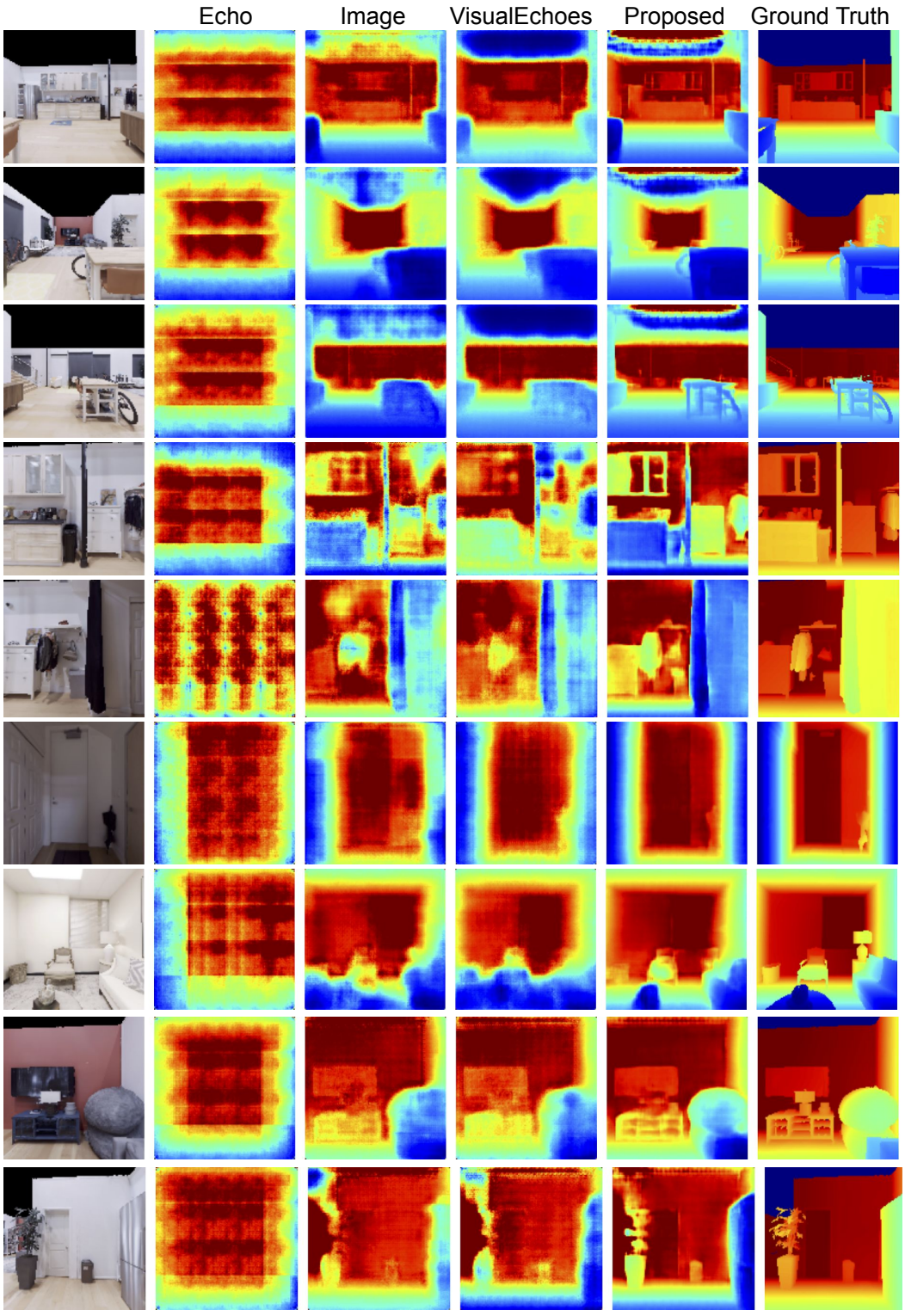

Figure 1. **Qualitative results for depth estimation on Replica dataset.** From left to right - input image, depth estimation using only echoes, depth estimation using only image, depth estimation from Visual Echoes, depth estimation using proposed method, ground truth depth map. The proposed method has better depth estimation in complicated scenes containing many objects causing frequent depth variations (e.g. row 1, row 4). It also provides robust depth estimation along boundaries of objects (e.g. rows 3,7,8). When the individual depth estimations from image and echo are poor leading to poor depth estimation (closer to image) using Visual Echoes, while the proposed method provides closer to ground truth estimation such as cabinets (row 4), door (row 9) which are completely missed by other methods.

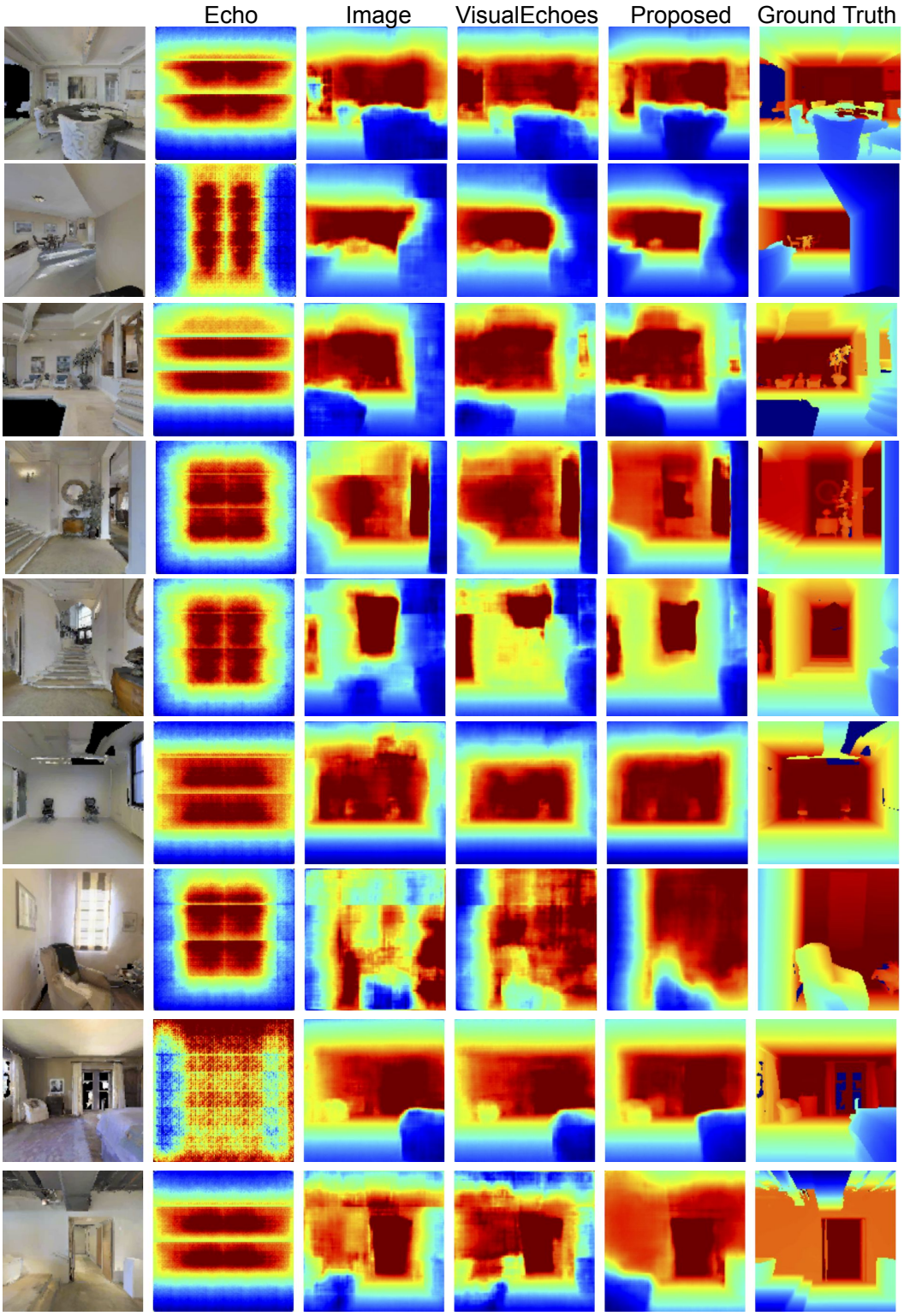

Figure 2. **Qualitative comparisons for depth estimation on Matterport3D dataset.** From left to right - input image, depth estimation using only echoes, depth estimation using only image, depth estimation from Visual Echoes, depth estimation using proposed method, ground truth depth map. We observe that the proposed method consistently provides better depth map estimation of smaller/farther objects (such as chairs cf. other methods in row 6) and also at object boundaries (rows 1,4,5). It also provides closer to ground truth results on illumination changes (row 7). We also observe that when image and echo depth estimations individually yield poor results, Visual Echoes tend to perform poorly as well while the proposed method is still able to estimate better depth (row 7).

Figure 3. **Visualization of attention maps on Replica dataset.** From left to right - input image, attention map from Echo Net, attention map from Visual Net.

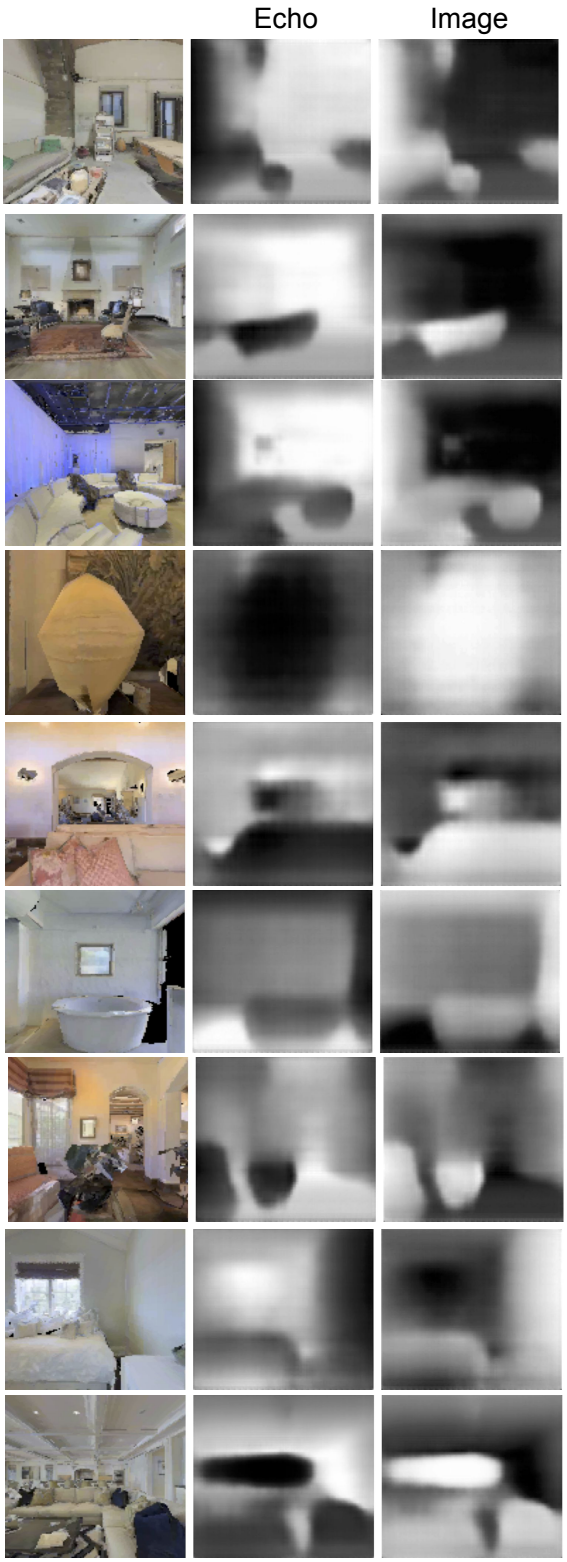

Figure 4. **Visualization of attention maps on Matterport3D dataset.** From left to right - input image, attention map from Echo Net, attention map from Visual Net.