Beyond Image to Depth: Improving Depth Prediction using Echoes

Kranti Kumar Parida¹ Siddharth Srivastava² Gaurav Sharma¹³
¹ IIT Kanpur ² CDAC Noida ³ TensorTour Inc.
{kranti, grv}@cse.iitk.ac.in, siddharthsrivastava@cdac.in

Abstract

We address the problem of estimating depth with multimodal audio visual data. Inspired by the ability of animals, such as bats and dolphins, to infer distance of objects with echolocation, some recent methods have utilized echoes for depth estimation. We propose an end-to-end deep learning based pipeline utilizing RGB images, binaural echoes and estimated material properties of various objects within a scene. We argue that the relation between image, echoes and depth, for different scene elements, is greatly influenced by the properties of those elements, and a method designed to leverage this information can lead to significantly improved depth estimation from audio visual inputs. We propose a novel multimodal fusion technique, which incorporates the material properties explicitly while combining audio (echoes) and visual modalities to predict the scene depth. We show empirically, with experiments on Replica dataset, that the proposed method obtains 28% improvement in RMSE compared to the state-of-the-art audio-visual depth prediction method. To demonstrate the effectiveness of our method on larger dataset, we report competitive performance on Matterport3D, proposing to use it as a multimodal depth prediction benchmark with echoes for the first time. We also analyse the proposed method with exhaustive ablation experiments and qualitative results. The code and models are available at https://krantiparida.github.io/projects/bimgdepth.html

1. Introduction

Humans perceive the surroundings using multiple sensory inputs such as sound, sight, smell and touch, with different tasks involving different combinations of such inputs. In computer vision, multimodal learning has also gained interest. As one popular stream, researchers have leveraged audio and visual inputs for addressing challenging problems. These problems can be broadly divided into three categories: (i) using audio modality only as the input, to learn a seemingly visual task, e.g. using echo for depth prediction [9], (ii) using visual modality as auxiliary information for an audio task, e.g. using videos to convert mono audio to binaural audio [14], and (iii) using both audio visual modalities together, e.g. for depth prediction [12]. Here, we follow the third line of work, and address the problem of depth map prediction using both audio and visual inputs. Studies in psychology and perception indicate that both sound and vision complement each other, i.e. visual information helps calibrate the auditory information [22] while auditory grouping helps solve visual ambiguity [42]. Many animals, like bats and dolphins, use echolocation to estimate the distances of objects from them. Visually impaired humans have also been reported to use echolocation [1]. Motivated by such cases, Christensen et. al [9, 10] recently showed that depth maps can be predicted directly from stereo sound. Gao et. al [12] showed that by fusing features from binaural echoes with the monocular image features, depth estimation can be improved. Inspired by these findings, we work with similar reasoning, i.e. sound contains useful information to predict depth, and that echoes, used along with monocular images, improve depth estimation.

Going beyond the current methods which do simple combinations of features from echoes and images [12], we
argue that the material properties of the objects in the scene significantly inform the spatial fidelity of the two streams. Some objects may lend better depth estimates with echoes, while some may prefer the visual modality more. Deriving from this motivation, we propose a novel end-to-end learnable network with a multimodal fusion module. This novel module incorporates material properties of the scene and fuses the two modalities with spatial attention maps indicating the fidelity of the respective modality for different spatial locations. The material properties are automatically estimated using a sub-network initialized with training on auxiliary data on materials. As the final depth prediction, the method fuses the depth maps produced by the audio and visual inputs, modulated by the predicted attention maps. Fig. 1 illustrates the difference with a real output of an existing approach and the proposed approach, showing qualitative improvements.

We demonstrate the advantages of the proposed method with experiments on Replica [37] and Matterport3D [7] datasets. We outperform the previous state-of-the-art on Replica dataset by $\sim 28\%$ RMSE. On Matterport3D, which is more complex and larger (5x) than Replica, we provide results on the multimodal depth prediction task for the first time, and compare the proposed method with existing approaches and challenging baselines. We also show that the proposed network can estimate better depth with low resolution images. This is important in practical systems working on depth estimation from monocular images, as sensors capturing echoes can be used along with cameras, to not only enhance the performance of existing setup but also suffer lesser degradation in depth prediction with the reduction in the quality of images. Further, we give ablation experiments to systematically evaluate the different aspects of the proposed method.

In summary, we make the following contributions:

- We propose a novel end-to-end learnable deep neural network to estimate depth from binaural audio and monocular images.
- We provide exhaustive quantitative and qualitative results on Replica and Matterport3D datasets. On Replica, we outperform the previous state-of-the-art by $\sim 28\%$. On Matterport3D, we provide results benchmarking existing methods. The proposed method achieves state-of-the-art performance, outperforming the existing best method on Matterport3D by $\sim 4\%$.
- We provide exhaustive ablation experiments on the design choices in the network, and validate our intuitions with representative qualitative results.

2. Related Works

Audio-visual learning. Recently there has been a surge in interest in audio-visual learning. In one line of work, the correspondence between both the modalities are used to learn the representation in each of the individual modality, in a self-supervised manner [2, 3, 19, 27, 29]. In [2, 3], the authors used an auxiliary task, of predicting whether the audio and image pair correspond to each other, to learn representations in each of the modalities. In [29], the authors predicted if the audio and the video clip are temporally synchronized to learn representations. In [19], the authors advanced a step further and localized sound generating objects in the image by leveraging the correspondence between audio and the image. In one of the recent approach, the authors in [27], have used the spatial correspondence between 360$^\circ$ video and audio. In another line of work, the integration of both audio and video modality was done to increase the performance. Recently a variety of task such audio source separation [49, 48, 11, 13, 15], zero-shot learning [30, 26], saliency prediction [40], audio spatialization [14, 28] have used the information from both audio and video modalities to improve the performance cf. using single modality only.

Depth estimation without echoes. The depth estimation methods span from only monocular image based methods to multi modal methods. Usually, the modalities are sparse depth maps, LiDAR point clouds, bird’s eye views, and normal maps. Monocular depth estimation methods involve utilizing a single RGB image to estimate dense depth [47, 6, 39]. Many methods directly utilize single image [16, 23, 32] or estimate an intermediate 3D representations such as point clouds [43, 45] and bird’s eye views [35] to estimate dense depth maps. A few other methods work on combining RGB with sparse depth maps, normal maps etc. [31, 25] to estimate dense depth maps.

Depth estimation with echoes. In [9, 10] depth of the scene was estimated using only echoes received from a single audio pulse. This approach completely ignored the visual modality while estimating the depth. On similar lines, the authors in [41] estimated the depth of scene directly from binaural audio of the object itself. They did not have the ground truth depth map, and instead used a vision network to predict the ground truth depth map. This method although used the direct audio from the object but the performance of the system was always upper bounded by the predicted depth map from visual input. In all of the methods above, the authors used one modality in isolation and did not fuse multi-modal information to improve the perfor-
mance of the system. In [12] the authors used echolocation as a pre-training task for learning a better visual representation. The authors also gave a case study, where they showed that simply concatenating the audio features with the visual features improves depth prediction. We explore the idea further, of improving depth prediction by adding binaural echoes to image input as well, and propose a novel multimodal fusion method which incorporates automatically estimated material properties in addition. We give the comparison with existing approach in Fig. 2.

3. Approach

We are interested in using echoes for predicting depth map of a scene. There is a spatio-temporal relationship between the received echoes and the depth of the scene, i.e. the echoes received at different instant of time directly relate to the different depths in the scene. Let \( x(t) \) be the original audio signal and \( y(t) \) be the echo response of the scene. Assuming we have \( d_i \) distinct depths of materials \( m_i \) in the scene (discretized and grouped for simplicity of discussion), the obtained echo response can be approximated with a summation of time delayed original signal, reflecting off the materials at different depths. The amplitudes of the delayed signals will depend upon the material they hit respectively. Considering only first order echoes, say every distinct \( d_i \) object contributes to a time delay of \( t_i \), and the corresponding material changes the amplitude of the signal by a factor of \( a_i \) on an average. The final response could then be approximated as

\[
y(t) = x(t) + \sum_{i=1}^{k} a_i x(t - t_i). \tag{1}
\]

With \( v_s \) denoting the speed of sound in the medium, the time delay can be directly associated with depths as \( t_i = \frac{2d_i}{v_s} \). Further, the amplitudes, \( a_i \) of each time-delayed signal, are dependent on the acoustic absorption and reflection coefficient of the material.

Hence, the goal of making the network learn depth from the received echo, is influenced by two factors, (i) the relationship between the echoes and spatial depth variation in the scene, and (ii) the different acoustic properties of the scene objects. We propose a carefully tailored attention mechanism (Sec. 3.6) between the image and audio modalities for addressing the spatial variation aspect. In addition we also propose to incorporate material property estimation (Sec. 3.4) as a proxy, to account for different properties of the scene elements which inform sound and light reflection and absorption, and hence inform the final depth prediction.

1The final response will include the original signal as well, as the sound emitter and recorder are both turned on together for a brief period of time.

3.1. Overall architecture

We show the block diagram of the proposed network in Fig. 3. The network consists of the following components: (i) echo subnetwork, (ii) visual subnetwork, (iii) material properties subnetwork, (iv) the multimodal fusion module and finally (v) attention prediction subnetwork. The echo and visual subnetworks consist of encoder-decoder pairs which estimate depth maps of the scenes independently. We input three feature maps coming from echo, visual and material property subnetworks respectively to the multimodal fusion network. The multimodal fusion module produces the fused features which we then feed to the attention network. We further use the attention network to predict two attention maps, one each for the two depth maps obtained from echo and visual decoder networks respectively. We then combine the individual depth maps using these attention maps to output the final depth map of the scene. We now give the details of the different components below. Please also see the supplementary document for detailed layer-wise architecture of the method.

3.2. Echo Net for Echo to Depth

The echo net is an encoder decoder network which predicts depth from binaural echo input. We convert the time-domain echo response into a frequency domain spectrogram representation, \( E \in \mathbb{R}^{2 \times P \times Q} \), where \( P \) is the number of discrete time steps and \( Q \) is the number of frequency bins. We input the spectrogram to the encoder part of the network to obtain the encoded representation \( f_e \in \mathbb{R}^N \) of the echo, which is also one of the inputs to the multimodal fusion module. We then reshape the encoded vector to \( N \times 1 \times 1 \) and feed it to a series of fractionally strided convolution layers to get the depth map, \( D_e \in \mathbb{R}^{W \times H} \), where \( W, H \) are the width and the height of the input image. While the upsampling happens here from an extreme \( 1 \times 1 \) feature, i.e. there is practically no spatial information (except what might get coded in the feature vector itself), such depth prediction from audio has been reported by earlier works also with fair success [12].

3.3. Visual Net for Image to Depth

The visual net is also an encoder decoder network which predicts depth from monocular RGB image. The architecture of this network is inspired from U-Net [33], and consists of regular convolutional and fractionally strided convolutional layers with skip connections between them. We give the image, \( I \in \mathbb{R}^{3 \times W \times H} \) as input to the network which predicts the depth map, \( D_v \in \mathbb{R}^{W \times H} \). We also use it to obtain the visual features from the intermediate layer (output of last conv layer) of the network denoted, \( f_i \in \mathbb{R}^{N \times w \times h} \). We use this feature as one of the inputs to the multimodal fusion module as well.

8270
3.4. Material Net for Material Properties

We use this network to extract the material properties of the objects present in the scene. We use a ResNet-18 architecture [18] and feed the RGB image, \( I \in \mathbb{R}^{3 \times W \times H} \) as the input. We obtain a feature map, \( f_m \in \mathbb{R}^{N \times w \times h} \) which encodes the material properties over the spatial locations in the scene image. This feature is the third input to the multimodal fusion module. We initialize the material network with pretraining on a large materials dataset [4] with classes such as fabric, brick, asphalt, wood, metal, and then train it to end with the rest of the network. We expect this initial material encoding capability of the network to be a proxy for encoding properties related to sound and light absorption and reflection, which affect depth prediction. Although the network evolves with the end to end training, the attention maps obtained qualitatively validate our assumptions (Sec. 4.5).

3.5. Multimodal Fusion Module

The multimodal fusion module combines features from the three sources discussed above, i.e., echo \( f_e \in \mathbb{R}^{N} \), visual \( f_v \in \mathbb{R}^{N \times w \times h} \) and material \( f_m \in \mathbb{R}^{N \times w \times h} \). Given the motivation, discussed in Sec. 1, that different object might give different depth prediction performances with audio or visual modalities respectively, the multimodal fusion module helps us combine the modalities to provide as input to the attention prediction network (Sec. 3.6).

We perform two bilinear transforms on the features to obtain two fusion maps, \( f_{img}^j \) and \( f_{mat}^j \), where \( j = 1, 2, ..., K \) is the number of output channels in the bilinear transformation,

\[
\begin{align*}
    f_{img}^j(p, q) &= f_e^T A_{img}^j f_e(p, q) + b_{img}^j, \forall p, q \quad (2) \\
    f_{mat}^j(p, q) &= f_e^T A_{mat}^j f_m(p, q) + b_{mat}^j, \forall p, q \quad (3)
\end{align*}
\]

where, \((p, q)\) indexes the spatial coordinates, \(A_{img}^j, A_{mat}^j\) are learnable weights of dimension \(N \times N\) and \(b_{img}^j, b_{mat}^j\) are scalar biases.

We finally concatenate the fusion maps \( f_{img}^j \in \mathbb{R}^{N \times w \times h} \) and \( f_{mat}^j \in \mathbb{R}^{N \times w \times h} \) along the first dimension to get the final fusion map \( f^* = \text{concat}(f_{img}, f_{mat}) \) to be fed into the per-pixel attention network.

3.6. Attention Network

The attention network is the final part of the network which we use to predict the per-pixel attention map given the concatenated fusion maps obtained in the previous step. The network consists of a series of fractionally strided convolutional layers with a final Sigmoid layer to normalize the values in the range \([0, 1]\).

The output of the network is an attention map \( \alpha \in \mathbb{R}^{1 \times W \times H} \). We use the attention map \( \alpha \) for weighting the echo predicted depth map \( D_e \) and \( 1 - \alpha \) for the image predicted depth map \( D_i \). The final depth map \( \hat{D} \) is thus,

\[
\hat{D} = \alpha \odot D_e + (1 - \alpha) \odot D_i \quad (4)
\]

where, \( \odot \) denotes pointwise multiplication.
3.7. Loss Function and Training

We train the network following [20], and use the logarithm of depth errors. The loss is given as,

\[
\mathcal{L}(\hat{D}, D) = \frac{1}{WH} \sum_{p=1}^{W} \sum_{q=1}^{H} \ln(1 + \|D(p, q) - \hat{D}(p, q)\|_1),
\]

where \(D\) is the ground truth depth map.

The full optimization problem is given by

\[
\theta^e_c, \theta^e_i, \theta^a, \theta^f, \theta^m = \arg \min_{\theta_e, \theta_i, \theta_a, \theta_f, \theta_m} \mathcal{L}(\hat{D}, D).
\]

where, \(\theta_c, \theta_i, \theta_a, \theta_f, \theta_m\) are the parameters for echo to depth network, image to depth network, attention network, fusion module and material property network respectively. We ignore the undefined regions in the ground truth depth maps, and therefore, such regions do not contribute to the learning. Adding smoothness constraints [23] can potentially further improve the quality of generated depth, however we obtain good results without using them here. We train the full network in an end to end fashion using standard backpropagation for neural networks.

4. Experiments

4.1. Implementation Details

Datasets. We perform experiments on Replica [37] and Matterport3D [7] datasets. Both the datasets contain indoor scenes. Replica has a total of 18 scenes covering hotels, apartments, rooms and offices. Matterport3D contains 90 scenes. On Replica, we follow [12] and use 15 scenes for training and 3 for testing. On Matterport3D, we use 77 scenes for evaluation, out of which 59, 10 and 8 scenes are used as train, validation and test respectively. We simulate echoes on these datasets by using the precomputed room impulse response (RIR) provided by [8] using 3D simulator Habitat [34] which takes into consideration the scene geometry and the materials present in the scene. We obtain the echoes by convolving input audio signal with the RIR. We use the material recognition dataset MINC [4] for pretraining the material net.

Network Architecture. We use the same architecture for echo encoder, and image to depth encoder and decoder (Visual Net) as [12], for fair comparison and demonstrating the effectiveness of the proposed material and attention networks. We use the first four convolutional layers of ResNet-18 for the material property network. We initialize them with pretraining on ImageNet and MINC dataset.

Input Representation. The input to Visual Net and Material Net is a 128 × 128 RGB image. For input to Echo Net, we use the spectrogram of 60ms echo signal. For training on Replica, we use a sampling frequency of 44.1 kHz and for Matterport3D, we use a sampling frequency of 16 kHz. We use Hanning window of length 64 and 32 to compute spectrogram for Replica and Matterport3D respectively. We use FFT size of 512 for both the cases.

Metrics. Following earlier works in depth estimation, we report results on root mean squared error (RMSE), mean relative error (REL), mean log10 error, and the percentage \(\delta_t\) of pixels with both the relative error and its inverse under threshold \(t\) where \(t \in \{1.25, 1.25^2, 1.25^3\}\).

Due to space constraints, we provide more details on the datasets, network architecture, parameter settings and evaluation metrics in the supplementary material.

4.2. Experiment Design

We design the experiments below to demonstrate the following points. (i) Using audio and visual modalities together improves performance over either of them. (ii) Using material properties in addition improves further. (iii) Among the different ways to combine the three, i.e. visual, audio and material properties, the proposed attention based fusion performs the best. We demonstrate the first two points with ablation experiments where we combine the inputs by simple concatenation, followed by a decoder to predict the depth map (Sec. 4.3 first part). Further, we demonstrate the third point by comparing combination methods and showing that attention based combination performs the best (Sec. 4.3 second part).

We then compare our full method with existing state of the art approaches (Sec. 4.4). We also show experiments on degrading resolution of image (Sec. 4.6).

4.3. Ablation Study

Combination of echo, image and material properties. We show the results of combining the three inputs with simple concatenation in Tab. 1. With only binaural echo as input, the RMSE is 0.995, which improves to 0.673 when image is added as well. When material property features are used with echo, the RMSE improves to 0.523 i.e. an improvement of \(\sim 47\%\) over echo only input and \(\sim 22\%\) over image+echo input.

Lastly, when image and material property features are concatenated with echo features (all), the RMSE value further improves to 0.491 i.e. \(\sim 50\%\) over echo only input and \(\sim 27\%\) over echo+image input.

These experiments validate that even with simple fusion, material properties improve the performance of the system. We attribute the improvement to our intuition that adding material properties explicitly allows the network to internally module the audio and visual features.

In the following we demonstrate the proposed explicit multimodal fusion followed by attention based weighting performs much better than the simple concatenation.
Impact of multimodal fusion and attention. We now validate the efficacy of our audio visual fusion method, which uses a multimodal fusion module to predict attention over the modalities to combine them. We compare the proposed fusion method, denoted \texttt{bilinear} with two alternatives, i.e., a simple concatenation of features denoted \texttt{concat}, and a dot product based fusion denoted \texttt{dot}. All these methods fuse the features and use them to estimate attention weights. We also compare by the fusion method of VisualEchoes, denoted \texttt{dot}, \texttt{concat} and \texttt{bilinear}. All these methods fuse the features and use them with a decoder to predict depth map, i.e. it has no attention based fusion.

We show the results in Tab. 2. We observe that \texttt{bilinear}, with an RMSE of 0.249, performs best among the compared methods, highlighting that the proposed fusion is better than the simple concatenation or dot product based fusion. We also observe that \texttt{concat} performs better than VisualEchoes i.e. 0.259 cf. 0.346 RMSE. This indicates that attention maps (which are present in \texttt{concat} but absent in VisualEchoes) are important for better performance.

Fig. 4 further shows the training loss (left) and validation RMSE (right) plots. We observe that VisualEchoes suffers from severe overfitting (much higher val RMSE), which is mitigated on adding the material features (i.e. \texttt{concat}). This further reinforces the hypothesis that material properties play an important role in depth prediction.

To conclude, we demonstrated from the ablation experiments that, (i) adding material properties explicitly is helpful for audio visual depth prediction, (ii) the proposed fusion strategy is better than simpler alternatives, and (iii) attention based combination of depth maps is better than simple concatenation as used in previous methods, e.g. VisualEchoes.

4.4. Comparison to state-of-the-art

Baselines. We compare on Replica and Matterport3D against VisualEchoes [12] and competitive baselines. The baseline methods are \texttt{AVERAGE}, \texttt{ECHO2DEPTH} and \texttt{RGB2DEPTH}. \texttt{AVERAGE} refers to average depth value of all the samples in training set. \texttt{ECHO2DEPTH} refers to the depth estimation using only Echo Net (Sec. 3.2) and \texttt{RGB2DEPTH} refers to depth estimation using only Visual Net (Sec. 3.3).

Comparison on Replica dataset. We report results in Tab. 3. The proposed method outperforms all the compared methods on all the metrics. Specifically, it outperforms VisualEchoes by \( \sim 28\% \) on RMSE. We also observe that while the improvement of VisualEchoes w.r.t. \texttt{RGB2DEPTH} is marginal (0.346 cf. 0.374 i.e. 7.4\%), the proposed method is able to achieve an improvement of \( \sim 33\% \) (0.249 cf. 0.374 RMSE). Both the methods perform significantly better than \texttt{AVERAGE} and \texttt{ECHO2DEPTH} baselines.

Comparison on Matterport3D dataset. We report results in Tab. 4. We outperform echo only (\texttt{ECHO2DEPTH}), image only (\texttt{RGB2DEPTH}) and \texttt{AVERAGE} baselines on all the five metrics. Our method also outperforms the existing VisualEchoes method by \( \sim 4\% \) on RMSE and on all the metric after training the method for 300 epochs. Further, better results on \( \delta \) indicate that the proposed method has lower pixel wise relative error cf. VisualEchoes which are manifested in form of better depth estimation around edges (Sec. 4.5).

Since Matterport3D is a popular benchmark for depth estimation, we also compare our method with the state-of-the-art methods on sparse to dense depth map estimation. These methods use sparse depth maps as inputs, while we have no explicit depth information in the inputs. We also use a slightly smaller subset of Matterport3D, i.e. 77 cf. 90 scenes for other. The results are shown in Tab. 5 where we obtain better results than four out of five compared methods. While the performances are not directly comparable, it supports the argument that echo can be a viable modality for estimating depth from RGB and can potentially provide additional information that are usually obtained from explicit 3D representations such as sparse depth maps.
We report the results for the baseline and existing methods directly from [12].

Table 3. Comparison with existing methods on Replica dataset.

<table>
<thead>
<tr>
<th>Method</th>
<th>RMSE ((\mu))</th>
<th>REL ((%))</th>
<th>log10 ((\mu))</th>
<th>log10 ((\mu))</th>
<th>log10 ((\mu))</th>
<th>log10 ((\mu))</th>
</tr>
</thead>
<tbody>
<tr>
<td>AVERAGE</td>
<td>1.050</td>
<td>0.701</td>
<td>0.233</td>
<td>0.253</td>
<td>0.385</td>
<td>0.509</td>
</tr>
<tr>
<td>Echo2Depth</td>
<td>0.713</td>
<td>0.347</td>
<td>0.134</td>
<td>0.148</td>
<td>0.580</td>
<td>0.722</td>
</tr>
<tr>
<td>RGB2Depth</td>
<td>0.374</td>
<td>0.202</td>
<td>0.076</td>
<td>0.088</td>
<td>0.749</td>
<td>0.883</td>
</tr>
<tr>
<td>VisualEchoes [12]</td>
<td>0.346</td>
<td>0.172</td>
<td>0.068</td>
<td>0.080</td>
<td>0.798</td>
<td>0.905</td>
</tr>
<tr>
<td>Proposed Method</td>
<td>0.249</td>
<td>0.118</td>
<td>0.046</td>
<td>0.069</td>
<td>0.860</td>
<td>0.943</td>
</tr>
</tbody>
</table>

We provide the attention maps as inputs, while the compared methods do not.

Table 4. Comparison with existing methods on Matterport3D dataset.

<table>
<thead>
<tr>
<th>Method</th>
<th>RMSE ((\mu))</th>
<th>REL ((%))</th>
<th>log10 ((\mu))</th>
<th>log10 ((\mu))</th>
<th>log10 ((\mu))</th>
<th>log10 ((\mu))</th>
</tr>
</thead>
<tbody>
<tr>
<td>AVERAGE</td>
<td>1.913</td>
<td>0.714</td>
<td>0.233</td>
<td>0.253</td>
<td>0.385</td>
<td>0.509</td>
</tr>
<tr>
<td>Echo2Depth</td>
<td>1.778</td>
<td>0.507</td>
<td>0.192</td>
<td>0.204</td>
<td>0.646</td>
<td>0.642</td>
</tr>
<tr>
<td>RGB2Depth</td>
<td>1.090</td>
<td>0.260</td>
<td>0.111</td>
<td>0.122</td>
<td>0.592</td>
<td>0.802</td>
</tr>
<tr>
<td>VisualEchoes [12]</td>
<td>0.998</td>
<td>0.193</td>
<td>0.083</td>
<td>0.095</td>
<td>0.711</td>
<td>0.878</td>
</tr>
<tr>
<td>Proposed Method</td>
<td>0.950</td>
<td>0.175</td>
<td>0.079</td>
<td>0.093</td>
<td>0.733</td>
<td>0.886</td>
</tr>
</tbody>
</table>

We give qualitative results of both (i) depth prediction, and (ii) attention maps in this section. We first provide few qualitative examples for attention map of echo and image respectively in Fig. 5. We provide the attention maps for two examples each from Replica and Matterport3D. We make a general observation from these results that the echo attention map closely looks into the far off regular regions but completely ignores the finer structures in them. It gives higher weight to the walls in the scenes but ignores any objects present on it (e.g. monitor on third example and wall painting on fourth). It also ignores the chairs and sofa which are present on relatively lower depth values.

The image attention map, being complementary to the echo attention, gives higher weights to nearby objects such as the sofa, chair etc., but ignores the far off wall in all the examples. This could be mainly due to two reasons (i) the echoes from the nearby regions are observed almost instantaneously which is hard to distinguish from the original signal (ii) most of the nearby objects (e.g. sofa) are made up of sound absorbing materials which do not reflect audio signals strongly. This results suggest that our network tries to leverage the best of both worlds, to get the final depth map.

We give a few examples of reconstructed depth using our approach and also compare it with an existing approach in Fig. 6. We observe that our approach is quite robust to fine grained structure in the scene cf. VisualEchoes [12]. In the first example, VisualEchoes misses the small chairs in the scene while our method gives an idea of the structure. Again the boundaries in the first example are not very clearly identified by VisualEchoes, but are estimated almost perfectly in our case. We also observe similar trends in other three examples as well. The results from the individual modalities (image and echo) are not satisfactory, but do capture the overall structure of the scene. These results suggest that the networks are not individually powerful but their performance improves significantly when we effectively combine the information from both of them. We encourage the readers to look at the supplementary material for more such qualitative results.

4.5. Qualitative Results

We give qualitative results of both (i) depth prediction, and (ii) attention maps in this section. We first provide few qualitative examples for attention map of echo and image respectively in Fig. 5. We provide the attention maps for two examples each from Replica and Matterport3D. We make a general observation from these results that the echo attention map mostly looks into the far off regular regions but completely ignores the finer structures in them. It gives higher weight to the walls in the scenes but ignores any objects present on it (e.g. monitor on third example and wall painting on fourth). It also ignores the chairs and sofa which are present on relatively lower depth values.

The image attention map, being complementary to the echo attention, gives higher weights to nearby objects such as the sofa, chair etc., but ignores the far off wall in all the examples. This could be mainly due to two reasons (i) the echoes from the nearby regions are observed almost instantaneously which is hard to distinguish from the original signal (ii) most of the nearby objects (e.g. sofa) are made up of sound absorbing materials which do not reflect audio signals strongly. This results suggest that our network tries to leverage the best of both worlds, to get the final depth map.

We give a few examples of reconstructed depth using our approach and also compare it with an existing approach in Fig. 6. We observe that our approach is quite robust to fine grained structure in the scene cf. VisualEchoes [12]. In the first example, VisualEchoes misses the small chairs in the scene while our method gives an idea of the structure. Again the boundaries in the first example are not very clearly identified by VisualEchoes, but are estimated almost perfectly in our case. We also observe similar trends in other three examples as well. The results from the individual modalities (image and echo) are not satisfactory, but do capture the overall structure of the scene. These results suggest that the networks are not individually powerful but their performance improves significantly when we effectively combine the information from both of them. We encourage the readers to look at the supplementary material for more such qualitative results.

4.6. Experiments with varying image resolutions

As discussed in Sec. 4.3, the network can better exploit the combination of echo and image due to attention based integration. This motivates us to study the effect of resolution degradation. This is similar to the case where human vision degrades and the brain learns to compensate and adapt based on the auditory input [5, 38]. We evaluate the proposed method by progressively degrading the image resolution. We give the results in Tab. 6 by gradually reducing the resolution to \(\frac{1}{2}\), \(\frac{1}{4}\), \(\frac{1}{8}\), and \(\frac{1}{16}\) times the original image.

We observe that the proposed approach is more robust to reduction in resolution of the images than the compared methods. The performance of image only method degrades significantly when the downscaling factor is \(\frac{1}{4}\) of the original image size, while the proposed method still performs better than image only method with original resolution i.e. 0.281 RMSE for proposed method at \(\frac{1}{4}\) scale cf. 0.374 with
We observe that the proposed approach is able to preserve fine structures, and has better depth estimation at the boundaries when compared to the existing approach. See supplementary material for more qualitative results.

Further, we observe that even with very high downscaling of $\frac{1}{32}$x, we obtain a better RMSE as compared to VisualEchoes (0.446 cf. 0.593). In fact, VisualEchoes performs worse than even the image only method. Similar observations can be made at $\frac{1}{16}$x. We can also observe that the rate of decrease in RMSE from 1x to $\frac{1}{4}$x is more in VisualEchoes cf. the image only and the proposed method. This further highlights the efficacy of the proposed method.

5. Conclusion

We presented a novel method for estimating depth by combining audio and visual modalities. We hypothesised that material properties play a significant role in the task, and proposed to use automatic material property estimation to predict spatial attention maps which modulate and combine the outputs of audio and image based depth prediction. We showed with quantitative experiments on challenging benchmark datasets, that (i) adding material properties explicitly improves the depth prediction over audio and visual prediction, (ii) having an attention mechanism based fusion method is better than other simpler existing approaches for audio visual fusion. We also demonstrated qualitatively that the attention maps focus on interpretable areas for the two modalities. While audio attention maps tended to ignore materials which would diffuse or absorb the audio wave, the image based attention included those areas. We also demonstrated qualitatively that the proposed method performs better than existing method, especially near the depth edges and brings out the finer structures in the scene.

We further showed experiments with reduced image resolution where our method degraded gracefully, while the compared methods loses performances significantly. We even compared our method with existing methods for sparse to dense depth prediction, and obtained encouraging competitive results, while not using sparse dense data as input for our method. We would like to explore such multimodal fusion with other modalities like sparse point clouds in the future to obtain even higher quality depth predictions. Further, geometric prior [36] can also be leveraged to improve the results.

In conclusion, we believe that using echo for depth prediction, especially in combination with other modalities is a promising direction, especially given the low cost and wide availability of audio sensors.

Acknowledgment. Kranti Kumar Parida gratefully acknowledges support from the Visvesvaraya fellowship.
References


