Stereo Depth Estimation with Echoes

Chenghao Zhang[1,2], Kun Tian[1,2], Bolin Ni[1,2], Gaofeng Meng[1,2,3⋆], Bin Fan[4], Zhaoxiang Zhang[1,2,3], and Chunhong Pan[1]

[1] NLPR, Institute of Automation, Chinese Academy of Sciences
[2] School of Artificial Intelligence, University of Chinese Academy of Sciences
[3] CAIR, HK Institute of Science and Innovation, Chinese Academy of Sciences
[4] University of Science and Technology Beijing

Abstract. Stereo depth estimation is particularly amenable to local textured regions while echoes have good depth estimations for global textureless regions, thus the two modalities complement each other. Motivated by the reciprocal relationship between both modalities, in this paper, we propose an end-to-end framework named *StereoEchoes* for stereo depth estimation with echoes. A Cross-modal Volume Refinement module is designed to transfer the complementary knowledge of the audio modality to the visual modality at feature level. A Relative Depth Uncertainty Estimation module is further proposed to yield pixel-wise confidence for multimodal depth fusion at output space. As there is no dataset for this new problem, we introduce two Stereo-Echo datasets named Stereo-Replica and Stereo-Matterport3D for the first time. Remarkably, we show empirically that our StereoEchoes, on Stereo-Replica and Stereo-Matterport3D, outperforms stereo depth estimation methods by 25%/13.8% RMSE, and surpasses the state-of-the-art audio-visual depth prediction method by 25.3%/42.3% RMSE.

Keywords: depth estimation, multimodal learning, cross-modal volume refinement, relative depth uncertainty estimation

1 Introduction

Recent years have witnessed exciting attempts to leverage audio visual multimodal learning for depth estimation [8,12,28]. For the visual modality, learning depth from stereo images is appropriate for textured regions though not accurate in textureless areas due to matching ambiguity. In contrast, in the audio modality, the echo has a good depth estimation for textureless regions in spite of large errors in local details. This suggests that the two modalities can complement each other, which is also reflected in psychology and perception that auditory grouping helps solve visual ambiguity [44] while visual information helps calibrate the auditory information [20].

Previous work in audio-visual depth estimation dates back to BatVision [8], which predicts depth directly from binaural echoes. The performance was improved by concatenating features of monocular images and echoes in [12]. Parida

⋆ Corresponding author. (E-mail:{chenghao.zhang,gfmeng}@nlpr.ia.ac.cn)

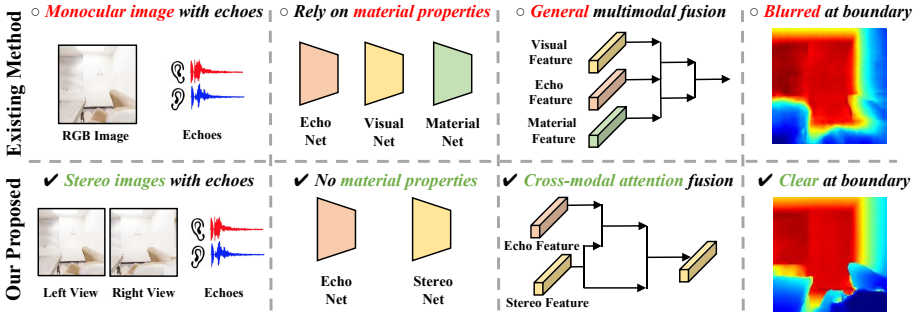

Fig. 1. Comparison of our method with the existing approach [28]. Our StereoEchoes learns depth from stereo images and echoes with no need for material properties estimation. Our method produces depth maps with clear object boundary.

et al. [28] explored the idea further by integrating a material properties estimation network with an attention mechanism, and achieved state-of-the-art performance. Unfortunately, material properties which rely on additional collected data with material annotations, are difficult to obtain for many environments. Besides, monocular depth estimation is an ill-posed problem thus the estimated depth maps are still blurry at local details and object boundaries.

To address these challenges, we argue that the above mentioned stereo depth estimation with echoes is a better configuration without demands on material properties. In general, the material of the object is usually reflected in the visual texture. Rich textures on objects are particularly amenable to stereo matching. From this point of view, stereo learning can be a good substitute for both monocular learning and material properties estimation. Although depth is not well estimated in textureless regions for stereo, the echo can play a complementary role for these areas. The multimodal learning of stereo images and echoes will make a dramatic leap on the performance of depth prediction.

Deriving from the above motivation, in this work, we propose an end-to-end framework named *StereoEchoes* for stereo depth estimation with echoes. To fully exploit the reciprocal relationship between the audio and visual modalities, we integrate both modalities at internal feature level and output space, respectively. At feature level, we propose a Cross-modal Volume Refinement module to transfer the complementary knowledge of echoes into the stereo features. On output space, we introduce a Relative Depth Uncertainty Estimation module to yield pixel-wise confidence for subsequent multimodal depth maps fusion. Our carefully designed cross-modal attention based fusion is empirically superior to previous general multimodal feature fusion. Fig. 1 shows the comparison of our method with the existing approach [28].

On the other hand, there are no specific datasets containing stereo images and echoes for depth estimation. Therefore, we introduce in this paper two StereoEcho datasets named Stereo-Replica and Stereo-Matterport3D from Replica [38]

and Matterport3D [3] respectively for multimodal stereo depth estimation benchmarks with echoes. The key point is to utilize the ground truth depth and given camera parameters to synthesize right-view images with original monocular images as the left-view images. The corresponding echoes are simulated using 3D simulators Habitat [33] and audio simulator SoundSpaces [5].

We evaluate the proposed StereoEchoes framework on the introduced Stereo-Echo datasets. We show in experiments that our method outperforms the state-of-the-art audio-visual depth prediction method [28] by 25.3% and 42.3% RMSE on Stereo-Replica and Stereo-Matterport3D. Compared with the challenging baselines that directly learn depth from the stereo, the improvements of our StereoEchoes are 25% and 13.8% respectively, demonstrating the superiority of incorporating echoes. Quantitative visualization shows that our method can produce more accurate depth maps on foreground objects and boundaries. Furthermore, extensive ablations validate the effectiveness of our methods.

Our contributions are summarized as follows:

- We propose to formulate the problem of stereo depth estimation with echoes, and introduce the StereoEchoes framework for this problem by utilizing the reciprocal relationship between the audio and visual modalities.
- Two modules of Cross-modal Volume Refinement and Relative Depth Uncertainty Estimation are designed for multimodal fusion at feature level and output space, which are superior to previous general fusion strategies.
- We further introduce two Stereo-Echo datasets named Stereo-Replica and Stereo-Matterport3D for evaluating the problem of multimodal stereo depth estimation with echoes.
- Experiments show that, on Stereo-Replica and Stereo-Matterport3D datasets, the proposed StereoEchoes outperforms the state-of-the-art audio-visual depth prediction method by 25.3% and 42.3% RMSE, and surpasses the challenging baseline of stereo depth estimation by 25% and 13.8%.

2 Related Work

Audio-visual Learning. Most videos contain both audio and visual information, which bridges the link between the two domains. The close connection between the two modalities has made a dramatic leap in audio-visual learning in recent years. In one line of work, the corresponding of the two modalities is used to learn the cross-modal representation in a self-supervised manner, which can be transferred well to a series of downstream tasks [2,26,50,1,6]. The representations can be learned by an auxiliary task of predicting the audio-visual correspondence [2], by predicting whether the audio and video frames are temporally synchronized [26] or spatially aligned [50], or by using cross-modal prediction for co-clustering [1]. One of the recent approaches further employs compositional contrastive learning to transfer audio-visual knowledge for better video representation learning [6]. In another line of work, the integration of both audio and visual modalities is exploited, which has promoted a variety of tasks such as audio source separation [54,11,13,15], audio spatialization [24,14],

audio-visual instance discrimination [23,25], saliency prediction [40], and action recognition [16]. Our work can be understood as exploring the complementary knowledge between audio and visual modality for the task of depth estimation.

Deep Stereo Matching. Depth is preferentially recovered by well-posed stereo matching due to its simple settings, high accuracy, and acceptable cost. Most deep stereo methods directly leverage rectified RGB stereo images to estimate disparity, which can be converted to depth with known camera baseline and focal length. One category of approaches utilizes correlation layers to encode matching information of left and right views [22,27,48], in which semantic [49] and edge information [36] are also incorporated as additional cues to improve performance. Another category of approaches focuses on building a 4D cost volume and leveraging 3D convolutions for cost aggregation to regress disparity or depth [19,4,17,52,7]. A few other methods work on combining other modalities to obtain accurate dense depth maps, like sparse LiDAR point clouds [30] or proxy disparity labels from a dedicated platform on-board the camera [31]. However, none of these techniques has been intended to cope with audio information, whereas this work demonstrates the ability of the audio modality to help predict better depth when combined with stereo images.

Multimodal Monocular Depth Estimation. Monocular depth estimation methods span from only monocular image based methods to multimodal methods. The former involves estimating dense depth maps from a single RGB image [53]. The latter usually increases other modalities including sparse depth maps [47], LiDAR point clouds [46,51], bird's eye views [37], and surface normal [32]. Recently, the audio modality is shown to be able to estimate the depth of the scene, like echoes from two ears [8] or binaural sounds of the object itself [41]. To fuse multimodal information from the audio and visual modalities to improve the performance, the authors in [12] leverage echolocation as a proxy task to learn better visual representations. They show that simply concatenating visual and audio features can yield better depth maps. This idea was further extended in [28] where a material properties estimation network is added and a multimodal fusion module is proposed with attention mechanisms for better depth estimation. Going beyond the ill-posed monocular depth estimation and inaccessible material properties, in this study, we propose to learn depth from stereo images using echoes with no requirements for material properties. Stereo learning on visual textures and the complementary knowledge from echoes can boost the performance of depth estimation to a further level.

3 Methods

3.1 Problem Formulation

In this paper, we focus on learning depth from stereo images with echoes. Given a pair of RGB stereo images, $(I_l, I_r) \in \mathbb{R}^{3\times W\times H}$, a Stereo Net is adopted to predict the disparity map, $D_v \in \mathbb{R}^{W\times H}$, where W, H are the width and height of images. With known camera baseline B and focal length f, the depth map Z_v

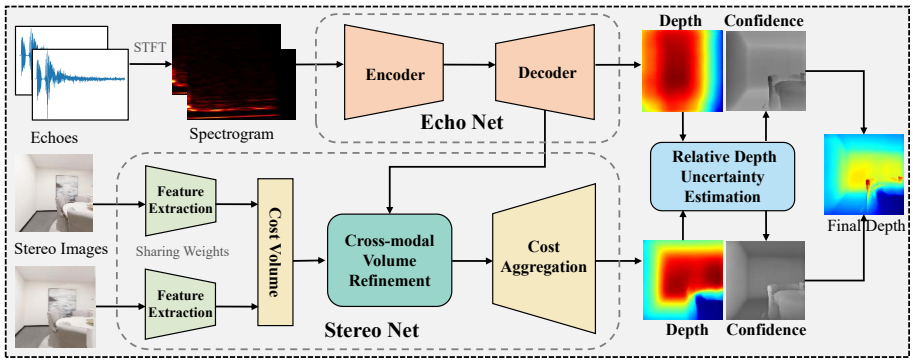

Fig. 2. Framework overview of our StereoEchoes. The multimodal data of stereo images and echoes pass through the Stereo Net and the Echo Net to yield their respective depth maps. The two networks interact at feature level through the Cross-modal Volume Refinement module. The final depth map is fused by the pixel-wise confidence produced by the Relative Depth Uncertainty Estimation module.

from the visual modality can be obtained by $Z_v = \frac{Bf}{D_v}$. The corresponding time-domain echo response is also given, which can be converted into a frequency 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. An Echo Net is employed to regress the depth map Z_a from the audio modality. The final depth map Z is obtained by fusing the depth maps predicted by both modalities.

3.2 Framework Overview

Fig. 2 depicts the framework of our method. As mentioned in Sec. 3.1, we take multimodal data of stereo images and echoes as input. The Echo Net and Stereo Net are adopted to yield the depth maps of the respective modalities. The Echo Net is an encoder-decoder network used in [12]. The Stereo Net inspired from [17] consists of feature extraction, cost volume construction, cost aggregation, and disparity regression. To fully exploit the reciprocal relationship between the audio and visual modalities, we integrate both modalities at internal feature level and output space, respectively. For feature-level integration, we propose a Cross-modal Volume Refinement module to transfer the complementary knowledge of echoes into the stereo cost volume for refinement. On the output space, we introduce a Relative Depth Uncertainty Estimation module to yield pixel-wise confidence of each modality. The final depth maps are obtained by fusing depth maps of both modalities with respective pixel-wise confidence.

3.3 Cross-modal Volume Refinement

Stereo matching approaches aim at learning structural information by comparing the similarity of local left and right patches to obtain the optimal disparity. As

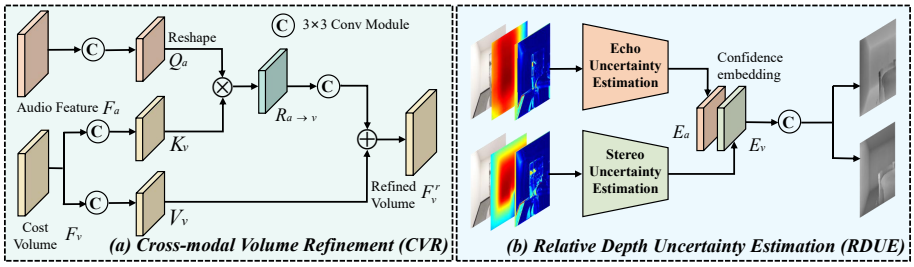

Fig. 3. Schematic diagram of Cross-modal Volume Refinement module (a) and Relative Depth Uncertainty Estimation module (b).

a result, stereo methods often succeed on richly-textured foreground objects but struggle to deal with textureless areas such as white walls. In contrast, depth prediction from echoes reported in earlier work [8,5,28] shows that the audio modality has a good estimation for textureless global regions like white walls despite large errors in the details. To fully exploit the reciprocal relationship between the two modalities, we propose to transfer the complementary knowledge of the audio modality to the visual modality at internal feature level.

Cost volume is the most significant internal feature in deep stereo networks, encoding all necessary information for succeeding disparity regression. The decoder features of the Echo Net also contain global characteristics related to depth regression. To this end, we design a Cross-modal Volume Refinement (CVR) module that utilizes echo features as a guide to help refine the cost volume of the Stereo Net. Fig. 3(a) shows the schematic diagram of CVR. Inspired by the cross-attention mechanism [42], we adopt audio features as query features and visual features as key-value features to learn audio-visual multimodal features for volume refinement.

Specifically, the CVR has two inputs: the audio feature $F_a \in \mathbb{R}^{B\times C_1\times H\times W}$ and the cost volume $F_v \in \mathbb{R}^{B\times C_2\times D\times H\times W}$. The convolutional modules with the kernel size 3×3 followed by BN and ReLU (W_Q^a, W_K^v, and W_V^v) are used to transform F_a and F_v to obtain the embeddings Q_a, K_v, and V_v:

$$Q_a = W_Q^a(F_a), K_v = W_K^v(F_v), V_v = W_V^v(F_v). \tag{1}$$

The audio-visual correlation $R_{a\to v}$ is computed by the Hadamard product $\circ$ between K_v and the reshaped $Q_a \in \mathbb{R}^{B\times(C_2D)\times H\times W}$ under the constraint of $C_1 = C_2 \times D$:

$$R_{a\to v} = \text{Reshape}(Q_a) \circ K_v. \tag{2}$$

The residual volume is learned from the multimodal correlation $R_{a\to v}$, which is further added to the input to yield the refined volume F_v^r as:

$$F_v^r = W_M(R_{a\to v}) + V_v, \tag{3}$$

where W_M denotes a convolutional module with the kernel size 3×3 followed by BN and ReLU.

In practice, we employ CVR at multi-scale audio and visual features, and the resolutions of feature maps of both modalities are aligned by upsampling or downsampling operations.

3.4 Relative Depth Uncertainty Estimation

As elucidated in Sec. 3.3, the audio modality estimates better depth in the global textureless regions, while the visual modality has a more accurate prediction on foreground objects. Multimodal depth maps can complement each other by fusion to obtain an optical depth map. The key lies in how to obtain the pixel-wise confidence of the respective depth map for each modality.

Deep neural networks provide a probability for each prediction, which is called epistemic uncertainty resulting from the model itself [18]. Earlier works employ Monte Carlo Dropout [10] to approximate the posterior distribution for uncertainty estimation, or ensemble [21]. For binocular vision, uncertainty estimation evolves into stereo confidence estimation. Various methods have been proposed for this that use the single-modal or bi-modal input [29,35,39]. These works mostly focus on the absolute confidence estimation of depth maps from the visual modality, but rarely explore the relative confidence estimation between two modalities. To this end, we propose a Relative Depth Uncertainty Estimation module (RDUE) to obtain the relative confidence of the visual modality compared to the audio modality for depth map fusion. Fig. 3(b) shows the architecture of RDUE.

Specifically, we first use the input left image I_l and the generated disparity map D to obtain the corresponding warped right image $\tilde{I}_r$ based on stereo reconstruction. Then, a small modal-specific fully-convolutional network (W_C^a and W_C^v) takes the concatenation of I_l, the respective depth map (Z_a and Z_v), and the corresponding pixel-wise error map (O_a and O_v) as inputs, which produces stereo confidence embeddings E_a and E_v. For both modalities, this can be expressed as:

$$\begin{aligned} E_a &= W_C^a(\mathrm{Cat}(I_l, Z_a, O_a)), \\ E_v &= W_C^v(\mathrm{Cat}(I_l, Z_v, O_v)), \end{aligned} \tag{4}$$

where the error map is calculated by the l_1 norm between the input right image and the warped one as $O = |I_r - \tilde{I}_r|$. Next, the relative depth confidence map $\mathcal{M}$ are learned from the stereo confidence embeddings of both modalities followed by a Sigmoid layer to normalize the values to $[0, 1]$:

$$\mathcal{M} = \mathrm{Sigmoid}(W_\circ(\mathrm{Cat}(E_a, E_v))), \tag{5}$$

where $W_\circ$ denotes a small fully-convolutional network for relative confidence estimation.

We use the relative confidence map $\mathcal{M}$ for weighting the depth map Z_v from stereo and $1 - \mathcal{M}$ for weighting the depth map Z_a from echoes. Thus the final depth map Z can be obtained by

$$Z = \mathcal{M} \odot Z_v + (1 - \mathcal{M}) \odot Z_a, \tag{6}$$

where $\odot$ denotes the element-wise product.

3.5 Objective Function

We train the Echo Net and Stereo Net jointly in an end-to-end manner following [28], and adopt the logarithm of depth errors as the loss function:

$$\mathcal{L}(Z, Z^*) = \ln(1 + ||Z - Z^*||_1), \tag{7}$$

where Z^* is the ground truth depth map.

4 Stereo-Echo Datasets

Lacking specific datasets for evaluating depth estimation from stereo images with echoes, we introduce two Stereo-Echo datasets named **Stereo-Replica** and **Stereo-Matterport3D** from two indoor visual scenes Replica [38] and Matterport3D [3], respectively. We describe the details of the visual scenes, echoes simulation, and stereo images synthesis as below.

Visual Scenes. Both Replica and Matterport3D datasets are rendered using open source 3D simulators, Habitat [33]. Replica has 18 scenes in total from 1740 images and 4 orientations $(90°, 180°, 270°, 360°)$, which covers hotels, apartments, rooms, and offices. Following [12], we use 15 scenes for training and 3 scenes for testing. On Matterport3D, we use 77 scenes of real-world homes from 16844 images and 4 orientations for evaluation following [28]. The training, validation, and testing sets consist of 59, 10, and 8 scenes, respectively.

Echoes Simulation. We use the audio simulator SoundSpaces [5] for realistic echoes simulation. The visual scene of the respective dataset is firstly divided into grids along navigation points. We place the source and receiver at the same point for sending the audio signal and receiving the echoes, respectively. Following [12], a 3 ms sweep signal is adopted as the source audio spanning the human hearing range (20Hz to 20kHz). The Room Impulse Response (RIR) is then calculated in four orientations using audio ray tracing [43]. Finally, we obtain the echoes by convolving the input audio signal with the RIR. Here, the sampling rates of the source and received echoes are 44.1 kHz and 16 kHz for Replica and Matterport3D, respectively. We encourage the interested readers to refer to [12] for more details.

Stereo Images Synthesis. We utilize RGB images and their ground truth depth to generate stereo image pairs similar to [45]. To simulate camera baselines and focal lengths, we directly convert the depth Z to disparity D with $D = \frac{Z_{max}}{Z}$. The RGB image is regarded as the left image I_l and the right image I_r is synthesized via forward warping [34]. Specifically, we translate each pixel of left image D pixels to the left side and perform linear interpolation to obtain the right image. However, pixels in I_r with no matching pixels in I_l may manifest themselves as holes in the synthesized image I_r. To address this, following image painting techniques [9], we fill the missing regions with a texture from a randomly selected image from the same scene. In this way, the interference from black holes with depth prediction is mitigated. Fig. 4 presents several synthesized examples with corresponding disparity maps. The datasets are available at https://github.com/chzhang18/Stereo-Echo-Datasets.

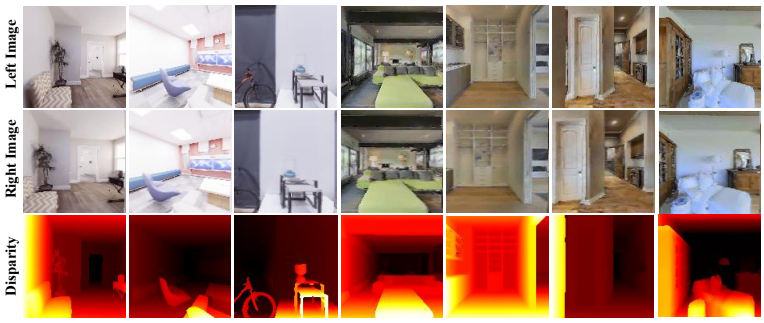

Fig. 4. Synthesized examples of stereo image pairs and disparity maps on Stereo-Replica dataset [38] (left three cases) and Stereo-Matterport3D dataset [3] (right four cases). Brighter colors indicate larger disparity values.

5 Experiments

In this section, we first introduce the implementation details and evaluation metrics. We then conduct experiments on our Stereo-Echo datasets, Stereo-Replica and Stereo-Matterport3D, to demonstrate the superiority of our method. Next, extensive ablation studies are provided to analyze the contribution of each proposed module in our framework, as well as the extensions for other stereo networks. Finally, we show the qualitative results to further validate the effectiveness of our method.

5.1 Experimental Setup

Our method is implemented with PyTorch. The input to the Stereo Net is 128×128 RGB stereo images. Color normalization is used for data preprocessing without any data augmentation. The maximum disparity is set to 32. For input to the Echo Net, following [28], 60ms echo signal is used to compute spectrogram with FFT size of 512. For Replica, a $2 \times 257 \times 166$ spectrogram is obtained using Hanning window of length 64 and hop length of 16. For Matterport3D, a $2 \times 257 \times 121$ spectrogram is produced using Hanning window of length 32 and hop length of 8. For network architecture, the Echo Net is used in [28] and the Stereo Net is adopted from [48]. Detailed architectures of the two networks and the proposed Cross-modal Volume Refinement and Relative Depth Uncertainty Estimation are provided in Supplementary Materials.

In training, we employ Adam ($\beta_1 = 0.9$, $\beta_2 = 0.999$) as the optimizer with learning rate of 0.001. The batch size is set to 50 on a single TITAN-RTX GPU. The entire model is trained from scratch. The training epoch is set to 100 for Replica and 50 for Matterport3D. The code is available at https://github.com/chzhang18/StereoEchoes.

Table 1. Evaluation results on the Stereo-Replica dataset. E refers to echoes, M refers to monocular images, and S refers to stereo images.

Methods	Modality	RMSE(↓)	REL(↓)	log10(↓)	$\delta_{1.25}$(↑)	$\delta_{1.25^2}$(↑)	$\delta_{1.25^3}$(↑)
ECHO2DEPTH	E	0.713	0.347	0.134	0.580	0.772	0.868
RGB2DEPTH	M	0.374	0.202	0.076	0.749	0.883	0.945
VisualEchoes [12]	$E+M$	0.346	0.172	0.068	0.798	0.905	0.950
BI2D [28]	$E+M$	0.249	0.118	0.046	0.869	0.943	0.970
STEREO2DEPTH	S	0.248	0.069	0.030	0.945	0.980	0.989
StereoEchoes	$\boldsymbol{E+S}$	**0.186**	**0.051**	**0.022**	**0.964**	**0.985**	**0.993**

Table 2. Evaluation results on the Stereo-Matterport3D dataset. E refers to echoes, M refers to monocular images, and S refers to stereo images.

Methods	Modality	RMSE(↓)	REL(↓)	log10(↓)	$\delta_{1.25}$(↑)	$\delta_{1.25^2}$(↑)	$\delta_{1.25^3}$(↑)
ECHO2DEPTH	E	1.778	0.507	0.192	0.464	0.642	0.759
RGB2DEPTH	M	1.090	0.260	0.111	0.592	0.802	0.910
VisualEchoes [12]	$E+M$	0.998	0.193	0.083	0.711	0.878	0.945
BI2D [28]	$E+M$	0.950	0.175	0.079	0.733	0.886	0.948
STEREO2DEPTH	S	0.636	0.058	0.026	0.943	0.979	0.990
StereoEchoes	$\boldsymbol{E+S}$	**0.548**	**0.049**	**0.021**	**0.958**	**0.984**	**0.992**

5.2 Comparison with State-of-the-art Methods

We compare on Stereo-Replica and Stereo-Matterport3D datasets with state-of-the-art methods: VisualEchoes [12] and Beyond-Image-to-Depth [28] (mark it as BI2D for convenience). We also compare with three competitive baselines: ECHO2DEPTH that predicts depth only from Echo Net, STEREO2DEPTH that predicts depth only from Stereo Net, and RGB2DEPTH that predicts depth only from monocular images.

Table. 1 shows the comparison results with the above methods on the Stereo-Replica dataset. Our proposed method outperforms all the compared methods on all the metrics. We observe that STEREO2DEPTH achieves comparable performance to BI2D on RMSE, and surpasses it on other metrics benefiting from the advantage of binocular vision. When adding the audio modality, our StereoEchoes outperforms BI2D by 25.3% (0.186 cf. 0.249 RMSE). It is also worth noting that our StereoEchoes achieves an order of magnitude lower than VisualEchoes by 50.3% (0.186 cf. 0.374 RMSE).

Table. 2 shows the comparison results on the Stereo-Matterport3D dataset. Our StereoEchoes achieves the best performance, notably outperforming BI2D by 42.3% on RMSE. Although the performance of STEREO2DEPTH is quite remarkable, our StereoEchoes further improves it by 13.8% on RMSE. This is surprisingly higher than the improvement of BI2D w.r.t. RGB2DEPTH (0.950 cf. 1.090 i.e. 12.8%). This validates our claim that the configuration of stereo inputs with echoes is better than that of monocular inputs with echoes. Stereo depth prediction with echoes can fully utilize the complementary knowledge to achieve higher performance without the help of the material network.

Table 3. Ablation study of key components on the Stereo-Replica and Stereo-Matterport3D datasets.

Datasets	Methods	CVR	RDUE	RMSE(↓)	REL(↓)	log10(↓)	$\delta_{1.25}$(↑)	$\delta_{1.25^2}$(↑)	$\delta_{1.25^3}$(↑)
Replica	STEREO2DEPTH			0.248	0.069	0.030	0.945	0.980	0.989
	Ours w/o RDUE	✓		0.202	0.054	0.024	0.958	0.983	0.991
	Ours w/o CVR		✓	0.193	0.057	0.023	0.958	0.984	0.991
	Ours (full)	✓	✓	**0.186**	**0.051**	**0.022**	**0.964**	**0.985**	**0.993**
Mp3D	STEREO2DEPTH			0.636	0.058	0.026	0.943	0.979	0.990
	Ours w/o RDUE	✓		0.593	0.054	0.023	0.951	0.980	0.990
	Ours w/o CVR		✓	0.599	0.050	0.022	0.951	0.980	0.990
	Ours (full)	✓	✓	**0.548**	**0.049**	**0.021**	**0.958**	**0.984**	**0.992**

Table 4. Performance of different feature fusion strategies on Stereo-Replica.

Methods	RMSE(↓)	REL(↓)	log10(↓)	$\delta_{1.25}$(↑)	$\delta_{1.25^2}$(↑)	$\delta_{1.25^3}$(↑)
BI2D [28]	0.249	0.118	0.046	0.869	0.943	0.970
Concat	0.228	0.065	0.028	0.953	**0.983**	**0.991**
Bilinear	0.234	0.065	0.029	0.952	**0.983**	**0.991**
CVR	**0.202**	**0.054**	**0.024**	**0.958**	**0.983**	**0.991**

5.3 Ablation Study

In this section, we conduct detailed ablation studies to demonstrate the following points. (i) Using stereo images and echoes together with either CVR or RDUE improves the performance over only using stereo images. (ii) Among different fusion strategies at feature level or output space, our proposed method achieves the best performance. (iii) Our method can be embedded into other stereo networks.

Effectiveness of Key Components. We adopt STEREO2DEPTH as the ablation baseline to show the impact of CVR and RDUE as shown in Table 3. As can be seen, applying the Cross-modal Volume Refinement can significantly reduce the RMSE, *e.g.*, 18.5% on Stereo-Replica and 6.8% on Stereo-Mp3D. We conjecture that since the visual scenes in Replica have more textureless regions, the echo guidance is more effective thus the improvement on Replica is more. The δ metrics are also improved indicating that the complementary knowledge from the audio modality also helps reduce the pixel-wise relative errors.

In addition, compared with the baselines, RMSE is reduced by 22.2% and 5.8% on Stereo-Replica and Stereo-Matterport3D respectively by integrating the proposed Relative Depth Uncertainty Estimation. The results validate that the depth fusion is able to fully exploit the reciprocal relationship between the audio and visual modalities on textureless and textured areas through the pixel-wise confidence map.

Furthermore, using both CVR and RDUE can further reduce the RMSE by $3.6\% \sim 8.5\%$ on the two datasets, though the performance is already encouraging by adding either of them. As a result, our full method significantly outperforms the corresponding baselines on both datasets, especially an improvement of 25% RMSE on the Stereo-Replica dataset.

Table 5. Performance of different depth fusion strategies at output space on the Stereo-Replica dataset.

Methods	RMSE(↓)	REL(↓)	log10(↓)	$\delta_{1.25}$(↑)	$\delta_{1.25^2}$(↑)	$\delta_{1.25^3}$(↑)
Average	0.281	0.069	0.030	0.941	0.978	0.989
Linear Weighting	0.246	0.066	0.029	0.946	0.980	0.990
Echo Uncertainty	0.242	0.066	0.029	0.948	0.981	0.990
Stereo Uncertainty	0.234	0.065	0.029	0.950	0.982	0.990
Relative Uncertainty	**0.193**	**0.057**	**0.023**	**0.958**	**0.984**	**0.991**

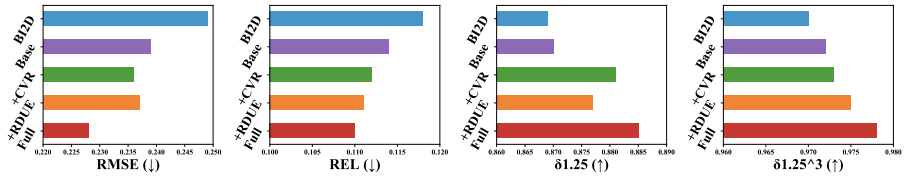

Fig. 5. Performance improvements by gradually adding proposed modules to AANet [48] on the Stereo-Replica dataset. The tag "full" refers to Base+CVR+RDUE.

Different Fusion Strategies. To further demonstrate the effectiveness of each module, we perform exhaustive comparisons with other alternative methods respectively. In Table 4, for the feature level fusion, two strategies are chosen for comparison, which are *concat* of audio and stereo features and *bilinear* transformation used in [28] with an attention network. We observe that our CVR, with RMSE of 0.202, achieves the best performance among the alternative strategies. Note that *concat* and *bilinear* perform better than BI2D benefiting from stereo settings, but are still inferior to our CVR. This highlights that our elaborately designed CVR for stereo and echo feature fusion is better than generic multimodal features fusion with attention.

In Table 5, for the fusion at output space, we take the average value and linear weighting [5] of the estimated depth from two modalities as baselines. We also compare with single-modal uncertainty-aware depth fusion strategies that leverage only audio or visual modality. One can observe that our method significantly outperforms these four compared strategies, indicating that multimodal relative uncertainty estimation is able to learn better pixel-wise confidence than single-modal absolute uncertainty estimation for depth fusion.

Stereo Backbone Network. We further extend our method to correlation-based stereo networks, *e.g.*, the lightweight stereo network AANet [48]. Since the aggregation network in AANet employs 2D convolutions, our CVR can adapt to the 3D matching volume while the RDUE remains constant. We reimplement all models with the training protocols detailed in Sec. 5.1. Fig. 5 depicts the results. It can be seen that our method delivers better performance using different stereo networks. This suggests that our designed method generalizes well to various stereo networks to improve their performance with echoes.

[5] The depth maps from stereo images and echoes are fused using weights of 0.9:0.1.

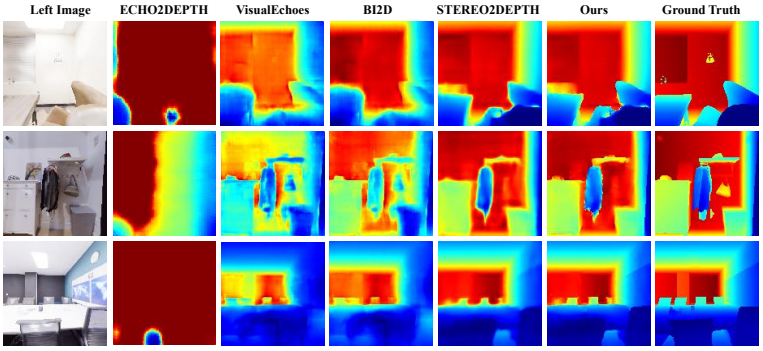

Fig. 6. Qualitative comparisons of different methods on Stereo-Replica. Our method produces better depth maps with fine structures and clear object boundaries.

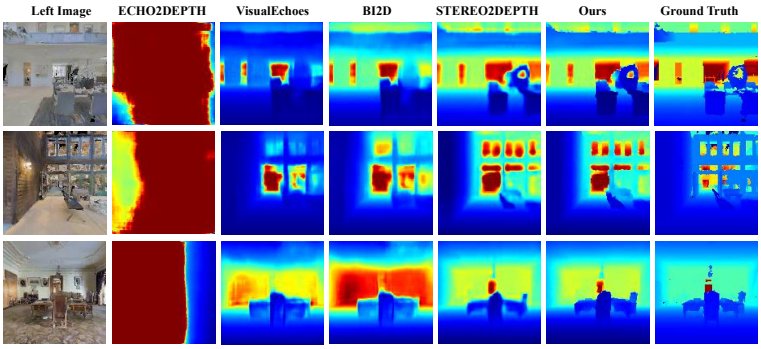

Fig. 7. Qualitative comparisons on Stereo-Matterport3D. Our method produces more accurate depth maps on both foregrounds and backgrounds.

5.4 Qualitative Results and Analysis

In addition to quantitative comparisons, we further provide qualitative visual analysis to illustrate the superiority of our method. Fig. 6 visualizes the depth map comparison with competing methods on the Stereo-Replica dataset. Compared to VisualEchoes and BI2D, STEREO2DEPTH is able to generate more accurate depth maps for foreground objects, such as chairs, benches, and clothes. This is mainly due to the rich texture and large disparity of foreground objects that are suitable for stereo learning. When echoes are integrated, the complementarity of both modalities makes the depth boundary sharper, since the audio modality is conducive to depth estimation for textureless backgrounds. Fig. 7 further shows the comparison results on the Stereo-Matterport3D dataset. Our method produces more accurate depth maps on both foreground objects and backgrounds. The reasons are two-fold. Firstly, the scenes in Matterport3D are rich in textures thus enhancing the stereo performance. Secondly, complementary knowledge of echoes further improves the depth maps of background objects that are far away, such as windows and murals on walls.

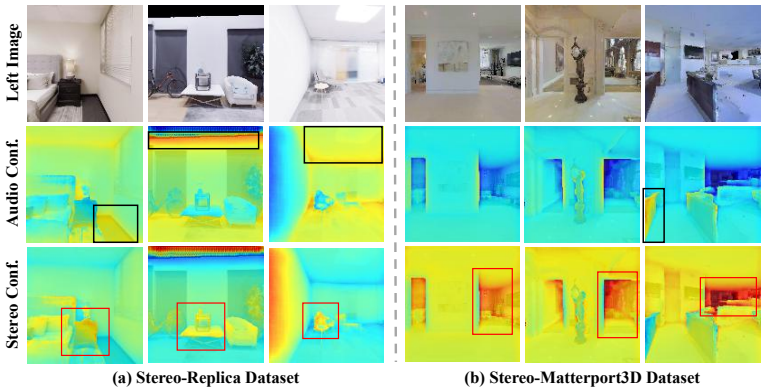

Fig. 8. Confidence map visualization on Stereo-Replica (left) and Stereo-Matterport3D (right). The echo modality produces high confidence for textureless areas (black box) whereas the visual modality attends more on richly-textured regions (red box). Warm colors represent high confidence while cool colors represent low confidence.

We further visualize the confidence maps for Echo Net and Stereo Net in Fig. 8. On Stereo-Replica, the audio modality is generally more confident in textureless regions (*e.g.*, white walls in the backgrounds) while the visual modality is more confident in textured foreground objects (*e.g.*, sofa and table). On Stereo-Matterport3D, the visual modality produces higher confidence in most regions, since most scenes have rich textures suitable for matching. Unfortunately, visual modalities tend to exhibit poor confidence in dark textureless regions, where audio modalities are mainly relied upon. An example is marked using a black box in the rightmost case in Fig. 8. The above analysis suggests that the audio and visual modalities can complement each other. Our method tries to leverage the best of the strengths of both modalities to yield the final depth.

6 Conclusion and Future Work

In this paper, we propose a new problem of predicting depth with stereo images and echoes and introduce the Stereo-Replica and Stereo-Matterport3D datasets as benchmarks. To exploit the reciprocal relationship of both modalities for addressing the problem, we have proposed the *StereoEchoes* framework consisting of the CVR module at the feature level and the RDUE module for multi-modal depth fusion. Extensive experiments on the two datasets validate that our method improves stereo depth estimation by adding echoes. In future work, we plan to extend our method to unsupervised conditions without ground truth depth and deploy our model on edge devices for robot navigation.

Acknowledgements This research was supported by the National Key Research and Development Program of China under Grant No.2018AAA0100400, and the National Natural Science Foundation of China under Grants 61976208, 62076242, 62071466, and the InnoHK project.

References

1. Alwassel, H., Mahajan, D., Korbar, B., Torresani, L., Ghanem, B., Tran, D.: Self-supervised learning by cross-modal audio-video clustering. In: NeurIPS. pp. 9758–9770 (2020) 3
2. Arandjelovic, R., Zisserman, A.: Look, listen and learn. In: ICCV. pp. 609–617 (2017) 3
3. Chang, A., Dai, A., Funkhouser, T., Halber, M., Niebner, M., Savva, M., Song, S., Zeng, A., Zhang, Y.: Matterport3d: Learning from rgb-d data in indoor environments. In: 3DV. pp. 667–676 (2017) 3, 8, 9
4. Chang, J.R., Chen, Y.S.: Pyramid stereo matching network. In: CVPR. pp. 5410–5418 (2018) 4
5. Chen, C., Jain, U., Schissler, C., Gari, S.V.A., Al-Halah, Z., Ithapu, V.K., Robinson, P., Grauman, K.: Soundspaces: Audio-visual navigation in 3d environments. In: ECCV. pp. 17–36 (2020) 3, 6, 8
6. Chen, Y., Xian, Y., Koepke, A., Shan, Y., Akata, Z.: Distilling audio-visual knowledge by compositional contrastive learning. In: CVPR. pp. 7016–7025 (2021) 3
7. Cheng, X., Zhong, Y., Harandi, M., Dai, Y., Chang, X., Drummond, T., Li, H., Ge, Z.: Hierarchical neural architecture search for deep stereo matching. In: NeurIPS. pp. 22158–22169 (2020) 4
8. Christensen, J.H., Hornauer, S., Stella, X.Y.: Batvision: Learning to see 3d spatial layout with two ears. In: ICRA. pp. 1581–1587 (2020) 1, 4, 6
9. Dwibedi, D., Misra, I., Hebert, M.: Cut, paste and learn: Surprisingly easy synthesis for instance detection. In: ICCV. pp. 1301–1310 (2017) 8
10. Gal, Y., Ghahramani, Z.: Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In: ICML. pp. 1050–1059 (2016) 7
11. Gan, C., Huang, D., Zhao, H., Tenenbaum, J.B., Torralba, A.: Music gesture for visual sound separation. In: CVPR. pp. 10478–10487 (2020) 3
12. Gao, R., Chen, C., Al-Halah, Z., Schissler, C., Grauman, K.: Visualechoes: Spatial image representation learning through echolocation. In: ECCV. pp. 658–676 (2020) 1, 4, 5, 8, 10
13. Gao, R., Feris, R., Grauman, K.: Learning to separate object sounds by watching unlabeled video. In: ECCV. pp. 35–53 (2018) 3
14. Gao, R., Grauman, K.: 2.5 d visual sound. In: CVPR. pp. 324–333 (2019) 3
15. Gao, R., Grauman, K.: Co-separating sounds of visual objects. In: ICCV. pp. 3879–3888 (2019) 3
16. Gao, R., Oh, T.H., Grauman, K., Torresani, L.: Listen to look: Action recognition by previewing audio. In: CVPR. pp. 10457–10467 (2020) 4
17. Guo, X., Yang, K., Yang, W., Wang, X., Li, H.: Group-wise correlation stereo network. In: CVPR. pp. 3273–3282 (2019) 4, 5
18. Kendall, A., Gal, Y.: What uncertainties do we need in bayesian deep learning for computer vision? In: NeurIPS (2017) 7
19. Kendall, A., Martirosyan, H., Dasgupta, S., Henry, P.: End-to-end learning of geometry and context for deep stereo regression. In: ICCV. pp. 66–75 (2017) 4
20. Kolarik, A.J., Moore, B.C., Zahorik, P., Cirstea, S., Pardhan, S.: Auditory distance perception in humans: a review of cues, development, neuronal bases, and effects of sensory loss. Attention, Perception, & Psychophysics **78**(2), 373–395 (2016) 1
21. Lakshminarayanan, B., Pritzel, A., Blundell, C.: Simple and scalable predictive uncertainty estimation using deep ensembles. In: NeurIPS (2017) 7

22. Liang, Z., Feng, Y., Guo, Y., Liu, H., Chen, W., Qiao, L., Zhou, L., Zhang, J.: Learning for disparity estimation through feature constancy. In: CVPR. pp. 2811–2820 (2018) 4
23. Morgado, P., Misra, I., Vasconcelos, N.: Robust audio-visual instance discrimination. In: CVPR. pp. 12934–12945 (2021) 4
24. Morgado, P., Nvasconcelos, N., Langlois, T., Wang, O.: Self-supervised generation of spatial audio for 360 video. In: NeurIPS (2018) 3
25. Morgado, P., Vasconcelos, N., Misra, I.: Audio-visual instance discrimination with cross-modal agreement. In: CVPR. pp. 12475–12486 (2021) 4
26. Owens, A., Efros, A.A.: Audio-visual scene analysis with self-supervised multisensory features. In: ECCV. pp. 631–648 (2018) 3
27. Pang, J., Sun, W., Ren, J., Yang, C., Yan, Q.: Cascade residual learning: A two-stage convolutional neural network for stereo matching. In: ICCV. pp. 878–886 (2017) 4
28. Parida, K.K., Srivastava, S., Sharma, G.: Beyond image to depth: Improving depth prediction using echoes. In: CVPR. pp. 8268–8277 (2021) 1, 2, 3, 4, 6, 8, 9, 10, 11, 12
29. Poggi, M., Mattoccia, S.: Learning from scratch a confidence measure. In: BMVC. vol. 2, p. 4 (2016) 7
30. Poggi, M., Pallotti, D., Tosi, F., Mattoccia, S.: Guided stereo matching. In: CVPR. pp. 979–988 (2019) 4
31. Poggi, M., Tonioni, A., Tosi, F., Mattoccia, S., Di Stefano, L.: Continual adaptation for deep stereo. IEEE Transactions on Pattern Analysis and Machine Intelligence (2021) 4
32. Qiu, J., Cui, Z., Zhang, Y., Zhang, X., Liu, S., Zeng, B., Pollefeys, M.: Deeplidar: Deep surface normal guided depth prediction for outdoor scene from sparse lidar data and single color image. In: CVPR. pp. 3313–3322 (2019) 4
33. Savva, M., Kadian, A., Maksymets, O., Zhao, Y., Wijmans, E., Jain, B., Straub, J., Liu, J., Koltun, V., Malik, J., et al.: Habitat: A platform for embodied ai research. In: ICCV. pp. 9339–9347 (2019) 3, 8
34. Schwarz, L.A.: Non-rigid registration using free-form deformations. Technische Universität München **6** (2007) 8
35. Shaked, A., Wolf, L.: Improved stereo matching with constant highway networks and reflective confidence learning. In: CVPR. pp. 4641–4650 (2017) 7
36. Song, X., Zhao, X., Fang, L., Hu, H., Yu, Y.: Edgestereo: An effective multi-task learning network for stereo matching and edge detection. International Journal of Computer Vision **128**(4), 910–930 (2020) 4
37. Srivastava, S., Jurie, F., Sharma, G.: Learning 2d to 3d lifting for object detection in 3d for autonomous vehicles. In: IROS. pp. 4504–4511 (2019) 4
38. Straub, J., Whelan, T., Ma, L., Chen, Y., Wijmans, E., Green, S., Engel, J.J., Mur-Artal, R., Ren, C., Verma, S., et al.: The replica dataset: A digital replica of indoor spaces. arXiv preprint arXiv:1906.05797 (2019) 2, 8, 9
39. Tosi, F., Poggi, M., Benincasa, A., Mattoccia, S.: Beyond local reasoning for stereo confidence estimation with deep learning. In: ECCV. pp. 319–334 (2018) 7
40. Tsiami, A., Koutras, P., Maragos, P.: Stavis: Spatio-temporal audiovisual saliency network. In: CVPR. pp. 4766–4776 (2020) 4
41. Vasudevan, A.B., Dai, D., Gool, L.V.: Semantic object prediction and spatial sound super-resolution with binaural sounds. In: ECCV. pp. 638–655 (2020) 4
42. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: NeurIPS (2017) 6

43. Veach, E., Guibas, L.: Bidirectional estimators for light transport. In: Photorealistic Rendering Techniques, pp. 145–167 (1995) 8
44. Watanabe, K., Shimojo, S.: When sound affects vision: effects of auditory grouping on visual motion perception. Psychological science **12**(2), 109–116 (2001) 1
45. Watson, J., Aodha, O.M., Turmukhambetov, D., Brostow, G.J., Firman, M.: Learning stereo from single images. In: ECCV. pp. 722–740 (2020) 8
46. Weng, X., Kitani, K.: Monocular 3d object detection with pseudo-lidar point cloud. In: ICCVW (2019) 4
47. Xiong, X., Xiong, H., Xian, K., Zhao, C., Cao, Z., Li, X.: Sparse-to-dense depth completion revisited: Sampling strategy and graph construction. In: ECCV. pp. 682–699 (2020) 4
48. Xu, H., Zhang, J.: Aanet: Adaptive aggregation network for efficient stereo matching. In: CVPR. pp. 1959–1968 (2020) 4, 9, 12
49. Yang, G., Zhao, H., Shi, J., Deng, Z., Jia, J.: Segstereo: Exploiting semantic information for disparity estimation. In: ECCV. pp. 660–676 (2018) 4
50. Yang, K., Russell, B., Salamon, J.: Telling left from right: Learning spatial correspondence of sight and sound. In: CVPR. pp. 9932–9941 (2020) 3
51. You, Y., Wang, Y., Chao, W.L., Garg, D., Pleiss, G., Hariharan, B., Campbell, M., Weinberger, K.Q.: Pseudo-lidar++: Accurate depth for 3d object detection in autonomous driving. In: ICLR (2019) 4
52. Zhang, F., Prisacariu, V., Yang, R., Torr, P.H.: Ga-net: Guided aggregation net for end-to-end stereo matching. In: CVPR. pp. 185–194 (2019) 4
53. Zhao, C., Sun, Q., Zhang, C., Tang, Y., Qian, F.: Monocular depth estimation based on deep learning: An overview. Science China Technological Sciences **63**(9), 1612–1627 (2020) 4
54. Zhao, H., Gan, C., Rouditchenko, A., Vondrick, C., McDermott, J., Torralba, A.: The sound of pixels. In: ECCV. pp. 570–586 (2018) 3