Stereo Depth Estimation with Echoes Supplementary Material

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

1 Network Architecture

We now provide the detailed architecture of each subnetwork of the proposed StereoEchoes.

Echo Net. The Echo Net is an encoder-decoder network used in [7]. The encoder network consists of 3 convolutional layers with the kernel size of 8×8, 4×4, 3×3, the stride of 4×4, 2×2, 1×1, and the number of output channels 32, 64, 8, respectively. Finally, a 1×1 convolutional layer is added to obtain a 512-dimensional echo feature vector. The decoder consists of 7 convolutional layers with the kernel size, stride, and padding of 4, 2, and 1, respectively. The corresponding number of output channels are 512, 256, 128, 64, 32, 16, and 1, respectively. Each layer is followed by BN and ReLU.

Stereo Net. The Stereo Net used in [3] consists of four parts, feature extraction, cost volume construction, cost aggregation, and disparity prediction. The detailed structure is listed in Table 1. For feature extraction, a ResNet-like siamese network with sharing weights is adopted to extract features of stereo images. The last feature maps of conv2, conv3, and conv4 are concatenated to construct the group-wise cost volume. Then, a 3D aggregation network is used to aggregate features from neighboring disparities and pixels to obtain refined cost volumes. It consists of a pre-hourglass module and three stacked 3D hourglass networks, which are connected to output modules to predict multi-scale disparity maps.

Cross-modal Volume Refinement. We adopt the output feature maps of layers 4, 5, and 6 of the decoder of Echo Net as the audio feature inputs of our CVR module. The three outputs of the Hourglass module of the Stereo Net are used as the visual feature inputs of CVR. We adopt 3×3 convolutions for downsampling or deconvolutions for upsampling to align the feature sizes of audio features and visual features. The 3D convolutional layers in CVR have the kernel size of $3 \times 3 \times 3$ with stride 1 and filter output size 32.

Relative Depth Uncertainty Estimation. The RDUE module mainly contains two parts, which are audio and visual uncertainty estimation networks.

⋆ Corresponding author.

Table 1. Detailed architecture of the Stereo Net. $[a \times a(\times a), c]$ refers to the 2D/3D convolution kernel size a and the output filter size c.

Name	Layer setting	Output dimension
Feature Extraction		
input		$H \times W \times 3$
conv0_x	$[3 \times 3, 32] \times 3$	$H/2 \times W/2 \times 32$
conv1_x	$[3 \times 3, 32] \times 3$	$H/2 \times W/2 \times 32$
conv2_x	$[3 \times 3, 64] \times 16$	$H/4 \times W/4 \times 64$
conv3_x	$[3 \times 3, 128] \times 3, dila = 1$	$H/4 \times W/4 \times 128$
conv4_x	$[3 \times 3, 128] \times 3, dila = 2$	$H/4 \times W/4 \times 128$
concat [conv2_16, conv3_3, conv4_3]		$H/4 \times W/4 \times 320$
Cost Volume		
Concat left and right feature		$D/4 \times H/4 \times W/4 \times 64$
Pre-hourglass		
3Dconv1_x	$[3 \times 3 \times 3, 32] \times 2$	$D/4 \times H/4 \times W/4 \times 32$
3Dconv2_x	$[3 \times 3 \times 3, 32] \times 2$	$D/4 \times H/4 \times W/4 \times 32$
pre-output	3Dconv1_2, 3Dconv2_2: Add	$D/4 \times H/4 \times W/4 \times 32$
Hourglass Module 1, 2, 3		
3Dconv3_x	$[3 \times 3 \times 3, 64] \times 2$	$D/8 \times H/8 \times W/8 \times 64$
3Dconv4_x	$[3 \times 3 \times 3, 128] \times 2$	$D/16 \times H/16 \times W/16 \times 128$
deconv1	deconv $[3 \times 3 \times 3, 64]$	$D/8 \times H/8 \times W/8 \times 64$
shortcut1	3Dconv3_2:$[3 \times 3 \times 3, 64]$	$D/8 \times H/8 \times W/8 \times 64$
plus1	deconv1, shortcut1:Add	$D/8 \times H/8 \times W/8 \times 64$
deconv0	deconv $[3 \times 3 \times 3, 32]$	$D/4 \times H/4 \times W/4 \times 32$
shortcut0	pre-output:$[1 \times 1 \times 1, 32]$	$D/4 \times H/4 \times W/4 \times 32$
hourglass-output	deconv0, shortcut0:Add	$D/4 \times H/4 \times W/4 \times 32$
Output Module 0, 1, 2, 3		
3Dconv4	$[3 \times 3 \times 3, 32]$	$D/4 \times H/4 \times W/4 \times 32$
3Dconv5	$[3 \times 3 \times 3, 1]$	$D/4 \times H/4 \times W/4 \times 1$
upsampling	bilinear interpolation	$D \times H \times W$
	disparity regression	$H \times W$

Both networks consist of 3 convolutional layers with the kernel size 3×3 and the number of output channels 32, 32, 16, respectively. The outputs of the two parts are concatenated to obtain the confidence embeddings, which are fed into 2 convolutional layers with the kernel size 3×3, 1×1, and output channels 32, 1 to yield the relative confidence map.

2 Analysis of Direct Depth Fusion

We are interested in investigating the impact of direct depth fusion from stereo images and echoes. To get some insights, we fuse the depth maps from both modalities (Z_v and Z_a) by linear weighting to obtain the final depth map:

$$Z = \alpha * Z_v + (1 - \alpha) * Z_a \tag{1}$$

where α is the depth fusion ratio. Fig. 1 shows the sensitivity analysis under different depth fusion ratios. We observe that utilizing average fusion ($\alpha = 0.5$) is inferior to directly using stereo images. This indicates that many errors exist in the depth map from echoes, which damages the stereo depth map. Notably, using a higher fusion ratio ($\alpha = 0.9$) can achieve better performance across all metrics, which validates the effectiveness of echoes in improving stereo depth estimation. Nevertheless, the improvement of direct depth fusion is limited since the depth fusion weight is global. By contrast, our designed Relative Depth

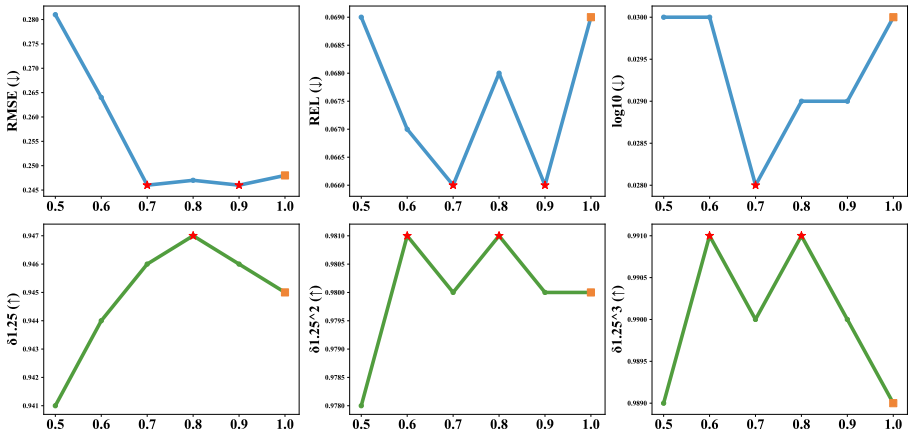

Fig. 1. Sensitivity analysis of depth fusion ratios α (horizontal axis) on the Stereo-Replica dataset. The top three metrics are RMSE, REL, and log10 (lower is better). The bottom three metrics are $\delta_{1.25}$, $\delta_{1.25^2}$, and $\delta_{1.25^3}$ (higher is better). Red stars mark the best fusion ratio. Orange squares mark the stereo baseline.

Uncertainty Estimation can predict modal-specific pixel-wise confidence thus greatly improving the performance with more fine-grained fusion.

Table 2. Comparison results with depth completion methods on Matterport3D. Red and blue represent the best and the second best results.

Methods	RMSE(↓)	MAE(↓)	$\delta_{1.25}$(↑)	$\delta_{1.25^2}$(↑)	$\delta_{1.25^3}$(↑)
Bilateral [9]	1.978	0.774	0.613	0.689	0.730
MRF [4]	1.675	0.618	0.651	0.780	0.856
Liu et al. [6]	1.653	0.610	0.663	0.792	0.861
Zhang et al. [11]	1.316	0.461	0.781	0.851	0.888
Huang et al. [5]	1.092	0.342	0.850	0.911	0.936
Senushkin et al. [8]	1.001	0.289	0.930	0.948	0.842
Srivastava et al. [10]	0.846	0.228	**0.954**	**0.985**	0.966
STEREO2DEPTH	**0.636**	**0.213**	0.943	0.979	**0.990**
StereoEchoes (Ours)	**0.548**	**0.198**	**0.958**	**0.984**	**0.992**

3 Comparison with Depth Completion Methods

Since Matterport3D is a popular benchmark for depth estimation, we also compare our method with the state-of-the-art depth completion methods. These methods can also be regarded as multimodal methods utilizing LiDAR point clouds or sparse depth maps as inputs to regress dense depth maps. The results are listed in Table 2. By comparison, we observe that STEREO2DEPTH

Table 3. Model parameters comparisons with Bi2D [7].

BI2D	StereoEchoes (ours)
Echo Net (8.30MB)	Echo Net (8.30MB)
Visual Net (16.66MB)	Stereo Net (6.91MB)
Material Net (11.69MB)	CVR module (0.26MB)
Attention Net (279.58MB)	RDUE module (0.04MB)
Total (316.23MB)	Total (15.51MB)

Table 4. Comparison of generalization ability for disparity estimation between training from scratch and StereoEcho pre-training.

Methods	EPE(↓)	D1-all(↓)	RMSE(↓)	$\delta_{1.25}$(↑)
Scratch	0.323	0.79%	12.353	0.876
StereoEchoes Pre-training	**0.308**	**0.53%**	**12.351**	**0.877**

obtains better RMSE and MAE although it is inferior to concurrent state-of-the-art method [10] on other metrics. This indicates that stereo learning can estimate more accurate absolute depth than monocular depth completion methods due to the well-posed settings and geometric completeness. When incorporating echoes, our method achieves the best performance on four out of five metrics. Notably, the RMSE significantly outperforms the other methods. This demonstrates that, by exploiting the reciprocal relationship between stereo images and echoes, better representations can be learned from the combination of the audio and visual modalities than from sparse depth maps.

4 Parameters Efficacy

To investigate the parameters efficiency of the proposed method, we list the parameters of each component in comparison with BI2D in Table 3. One can observe that our method employs a 20 times smaller model but achieves remarkable performance on the Stereo-Replica and Stereo-Matterport3D datasets as reported in main paper. This further validates our claim that, compared to monocular audio-visual depth estimation with the material network and heavy attention network, the combination of stereo images and echoes is a better configuration with lightweight parameters.

5 Generalizability on Real Data

We repurpose the learned stereo representation for disparity estimation on the real-world stereo dataset named InStereo2K [1] following VisualEchoes [2]. Table 4 shows the comparison results of pre-training on StereoEcho datasets and training from scratch on the test set. We adopt additional EPE and D1-all metrics to demonstrate the stereo performance. The stereo model initialized with the pre-trained StereoEchoes network achieves better stereo performance compared

to that trained from scratch. This suggests that stereo with echoes generalizes better than stereo only for disparity estimation.

6 Evaluation Metrics

Let Z_p and Z_p^* denote the predicted depth and ground truth depth for every valid pixel p. Here, valid pixels are those pixels with a total of N whose ground truth depth values are greater than zero. We adopt the following standard metrics: (1) Root mean square error (RMSE): $\sqrt{\frac{1}{N}\sum_p (Z_p - Z_p^*)^2}$. (2) Mean absolute relative error (REL): $\frac{1}{N}\sum_p \frac{|Z_p - Z_p^*|}{Z_p^*}$. (3) Mean $\log_{10}$ error (log10): $\frac{1}{N}\sum_p \left\|\log_{10} Z_p - \log_{10} Z_p^*\right\|$. (4) Accuracy under threshold t: $\max\left(\frac{Z_p^*}{Z_p}, \frac{Z_p}{Z_p^*}\right) = \delta < t \left(t \in \left[1.25, 1.25^2, 1.25^3\right]\right)$.

7 More Qualitative Results

We provide more qualitative results of audio-visual depth estimation using various approaches on the Stereo-Replica and Stereo-Matterport3D datasets in Fig. 2 and Fig. 3 respectively. The visualizations of the confidence maps from Echo Net and Stereo Net are shown in Fig. 4.

References

1. Bao, W., Wang, W., Xu, Y., Guo, Y., Hong, S., Zhang, X.: Instereo2k: A large real dataset for stereo matching in indoor scenes. Science China Information Sciences **63**(11), 1–11 (2020) 4
2. Gao, R., Chen, C., Al-Halah, Z., Schissler, C., Grauman, K.: Visualechoes: Spatial image representation learning through echolocation. In: ECCV. pp. 658–676 (2020) 4, 6, 7
3. Guo, X., Yang, K., Yang, W., Wang, X., Li, H.: Group-wise correlation stereo network. In: CVPR. pp. 3273–3282 (2019) 1
4. Harrison, A., Newman, P.: Image and sparse laser fusion for dense scene reconstruction. In: Field and Service Robotics. pp. 219–228 (2010) 3
5. Huang, Y.K., Wu, T.H., Liu, Y.C., Hsu, W.H.: Indoor depth completion with boundary consistency and self-attention. In: CVPRW (2019) 3
6. Liu, J., Gong, X.: Guided depth enhancement via anisotropic diffusion. In: Pacific-Rim conference on multimedia. pp. 408–417 (2013) 3
7. Parida, K.K., Srivastava, S., Sharma, G.: Beyond image to depth: Improving depth prediction using echoes. In: CVPR. pp. 8268–8277 (2021) 1, 4, 6, 7
8. Senushkin, D., Romanov, M., Belikov, I., Patakin, N., Konushin, A.: Decoder modulation for indoor depth completion. In: IROS. pp. 2181–2188 (2020) 3
9. Silberman, N., Hoiem, D., Kohli, P., Fergus, R.: Indoor segmentation and support inference from rgbd images. In: ECCV. pp. 746–760 (2012) 3
10. Srivastava, S., Sharma, G.: Self attention guided depth completion using rgb and sparse lidar point clouds. In: IROS. pp. 2643–2650 (2021) 3, 4
11. Zhang, Y., Funkhouser, T.: Deep depth completion of a single rgb-d image. In: CVPR. pp. 175–185 (2018) 3

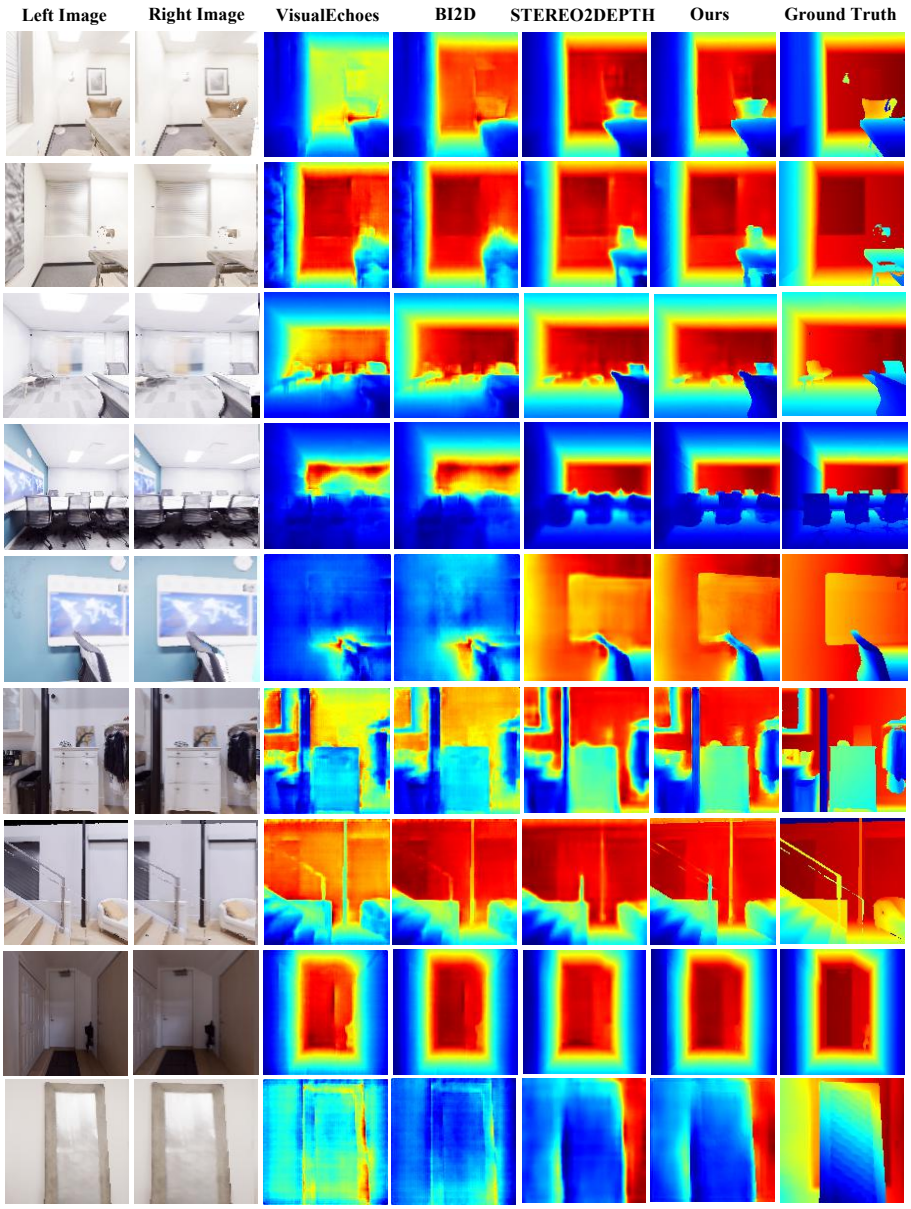

Fig. 2. Qualitative comparison results of depth maps on the Stereo-Replica dataset. From left to right are left image, right image, depth maps from VisualEchoes [2], BI2D [7], stereo images, our proposed StereoEchoes, and ground truth. Our method consistently produces more accurate depth maps with clear object boundaries (rows 1-5). It also produces closer results to ground truth for irregular objects (row 6), thin structures (row 7), and dark light (row 8). The performance in reflective areas needs to be further improved (row 9).

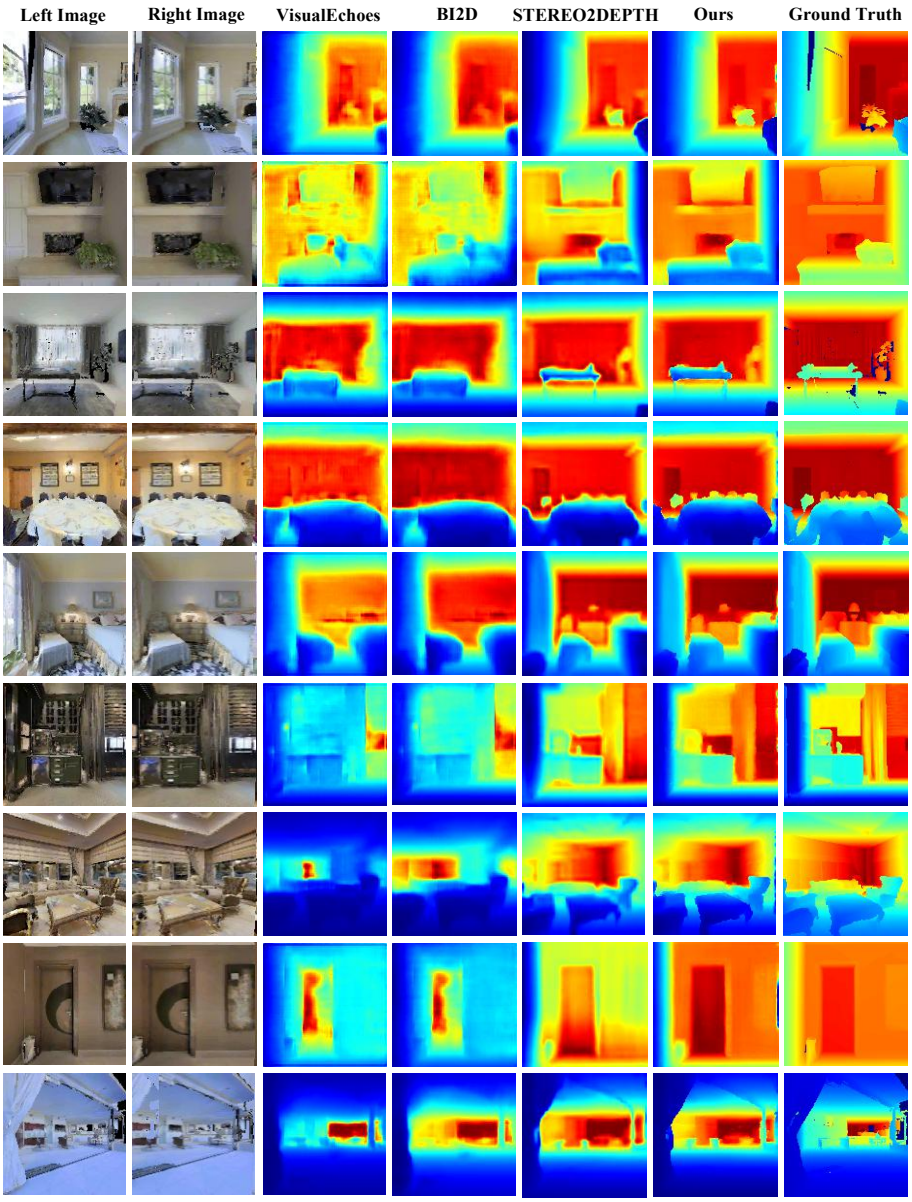

Fig. 3. Qualitative comparison results of depth maps on the Stereo-Matterport3D dataset. From left to right are left image, right image, depth maps from VisualEchoes [2], BI2D [7], stereo images, our proposed StereoEchoes, and ground truth. Since the visual scenes in Matterport3D are rich in texture that are beneficial for stereo learning, our method estimates more accurate depth maps for both foregrounds and backgrounds. The incorporation of echoes makes object boundaries clearer and sharper by the complementation of visual and audio modalities.

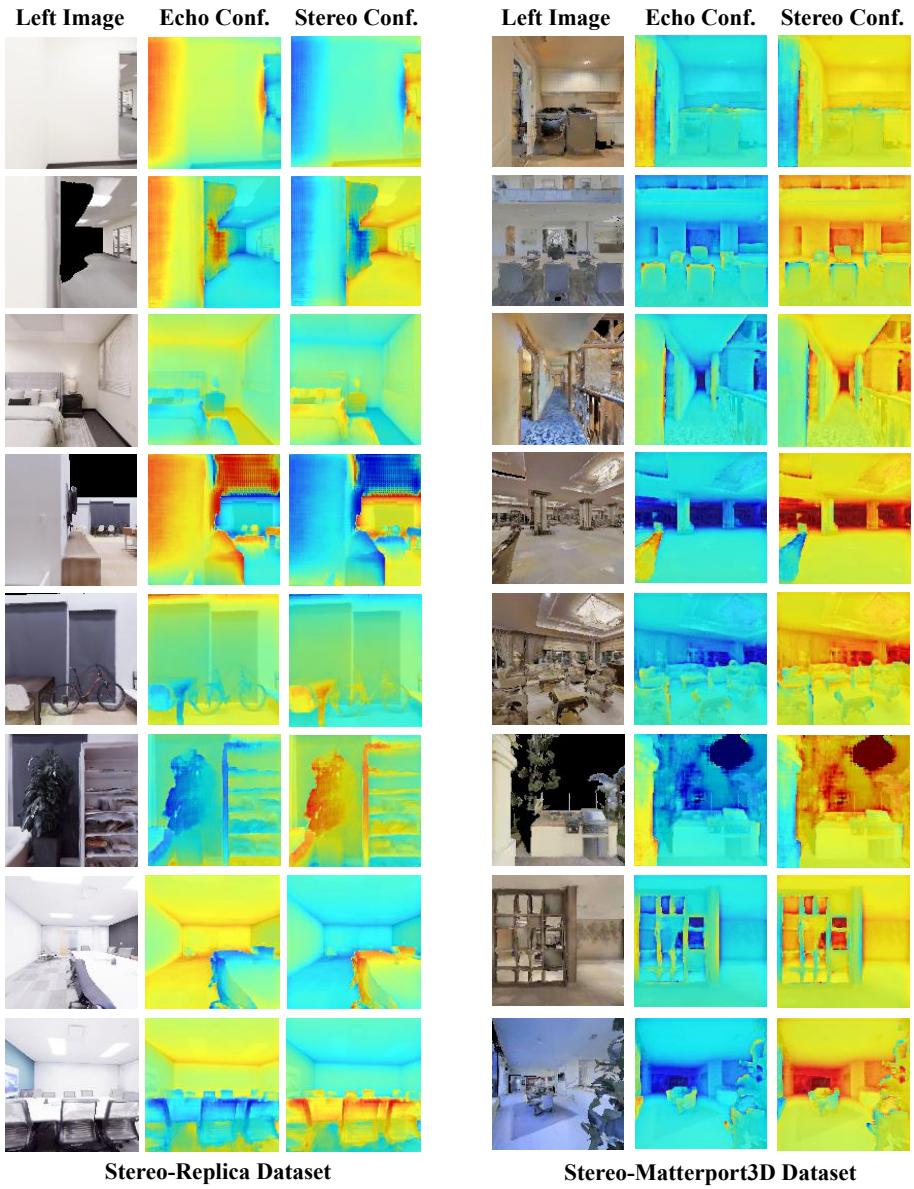

Fig. 4. Visualization of relative confidence maps from stereo images (Stereo Conf.) and echoes (Echo conf.) on the Stereo-Replica (left) and Stereo-Matterport3D (right) datasets.