AVS-Net: Audio-Visual Scale Net for Self-supervised Monocular Metric Depth Estimation

Xiaohu Liu, Sascha Hornauer, Fabien Moutarde, Jialiang Lu

AVS-Net: Audio-Visual Scale Net for Self-supervised Monocular Metric Depth Estimation

Xiaohu Liu[1] Sascha Hornauer[2] Fabien Moutarde[2] Jialiang Lu[1]
[1]SJTU Paris Elite Institute of Technology, Shanghai Jiao Tong University, Shanghai, China
[2]Center for Robotics, MINES Paris, PSL University, Paris, France

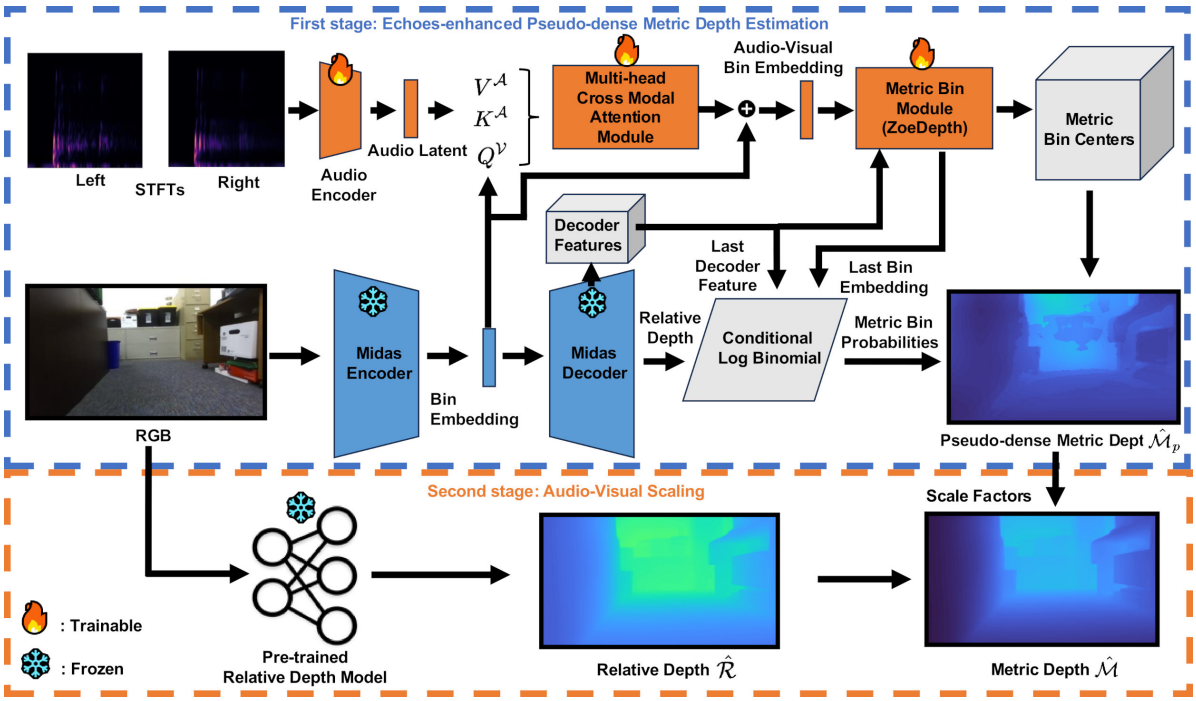

Figure 1. Illustration of the proposed two-stage method.

1. Introduction

Monocular depth estimation provides useful 3D information using simple hardware but is an ill-posed problem because the same RGB image can correspond to an infinite number of 3D scenes. Predictions fall into two categories: *Metric* depth estimation [1, 2, 6] returns the actual physical distance of pixels but tends to overfit and exhibits poor generalisation to novel datasets [2]. *Relative* depth estimation [5, 12, 11] predicts the relative distance between pixels and exhibits better generalisation across multiple datasets [9]. While relative prediction has its applications, absolute predictions on a physical metric scale are important in domains such as autonomous driving and robot navigation.

Integrating acoustic echoes into supervised depth estimation has shown robust improvement of visual depth quality and helped to resolve visual illusions [4, 8, 7, 3]. Echoes reflect from objects in the scene, travel at fixed speeds and can inform the network reliably of true distances. In the following, we present a method to exploit the time-of-flight information of echoes to transform relative into metric depth.

We showcase scale-corrections of self-supervised relative depth estimation [5] with a two-step method named **A**udio-**V**isual **S**cale**Net** (AVS-Net), to exploit the information in echoes for metric depth prediction. First we add audio to a recently proposed metric depth prediction method for improved supervised model training on an audio-visual dataset [2]. Then, scale factors are extracted to provide the missing scale information for relative depth models such as [5, 12, 11] or correct zero-shot models such as [2, 6, 10].

A similar approach [2] first pre-trains a model on 12 datasets with relative depth and then fine-tunes using metric depth. In contrast to learning one target depth distribution of a target metric dataset our proposed method first learns the relationship between echoes and depth. That relationship stays constant for arbitrary novel datasets. Once trained, our model can be used to scale-correct any other model trained on relative depth. That way, relative depth prediction methods can improve independently while our correction-model only re-scales the results informed by echoes. Our main contributions can be summarized as follows:

- We propose AVS-Net, a method using echoes to provide scale information for relative depth predictions.

- We scale-correct self-supervised relative-depth prediction and improve zero-shot metric-depth prediction.

- Evaluations on the real-world BatVision datasets [3] show our proposed Audio-Visual method significantly outperforms the visual-only counterpart by approximately 30% and 22% in δ_1 accuracy for the first-stage pseudo-dense[1] metric depth and the final scale-corrected metric depth map respectively. This demonstrates the effectiveness of echoes in providing valuable geometry scale information.

2. Methodology

Our approach is shown in Fig.1. Inspired by ZoeDepth [2], we furthur proposed and tested an echoes-enhanced two-stage method: In the first stage, Audio-Visual fusion of STFT (Short-Time Fourier Transform) representations of echoes is performed to better estimate metric bin centers and obtain enhanced pseudo-dense metric maps, denoted as $\hat{\mathcal{M}}_p$. In the second stage, the $\hat{\mathcal{M}}_p$ are used to obtain scale factors (medians for this paper) to combine with outputs of relative depth models, denoted as $\hat{\mathcal{R}}$, for metric depth estimation. The final scale corrected metric depth $\hat{\mathcal{M}}$ can be obtained by $\hat{\mathcal{M}} = \hat{\mathcal{R}} \times \texttt{MEDIAN}(\hat{\mathcal{M}}_p)/\texttt{MEDIAN}(\hat{\mathcal{R}})$.

AVS-Net uses a pre-trained Midas [9] encoder to obtain visual feature vectors. The echoes feature vector is obtained by adding an additional echoes embedding branch to encode ego-centric binaural echoes corresponding to the input RGB image. Audio-Visual fusion is performed by a cross-modal attention module [7]. These fused vectors are then passed through the *seed bin regressor* [2] to obtain the initial bin centres, which will be incrementally adjusted by the attraction layers proposed in ZoeDepth, and finally the metric bins are linearly combined by a log-binomial to obtain the pseudo-dense metric depth map.

The intuition of our approach is to decompose metric depth estimation into relative depth- and metric factor-estimation in order to combine the generalisation ability of relative depth models with scale information from echoes.

3. Experiments

We evaluate metric depth prediction of the AVS-Net and scale-correction of relative depth models, either trained self-supervised or zero-shot when trained on other datasets.

3.1. Dataset.

We evaluate on the recently released BatVision dataset [3], which has two parts, $BV1$ and $BV2$, recorded at different places. Both contain RGB-D pairs synchronized with recorded binaural echoes from emitted *Chirps* ranging from 20Hz to 20kHz. The two parts have very different average depth, being recorded in offices and in an historic university. We showcase the generalization between datasets by training the AVS-Net only on the training set of $BV2$, and the relative depth model on the training set of $BV1$.

For supervised training of AVS-Net, RGB-D and echoes are needed for training, validation and test. We follow the train-val-test split of $BV2$ [3], resulting in 1911 instances for training, 625 for validation and 584 for test. For each binaural audio signal, we follow original BatVision settings [3] to generate the STFT and resize it to the resolution of the RGB image before feeding it to the Audio Encoder.

For self-supervised training of relative depth models exploiting scene movement, consecutive frames are needed. Therefore we re-organize the original $BV1$ train- and validation instances to form consecutive frame triplets. We choose 20-frame time intervals for which we found suitable camera movements. For testing only single RGB-D images and echoes are needed. This results in 39165 train-triplets, 7378 validation-triplets and 4960 RGB-D-Echo instances for testing with the same settings for STFTs as for $BV2$. Self-supervised trained models are initialized with ImageNet weights.

Table 1. First stage pseudo-dense metric depth estimation results for AVS-Net. AV denotes Audio-Visual, V denotes Visual only.

Method	Input	Abs rel ↓	sq rel ↓	RMSE ↓	RMSE log ↓	δ_1 ↑	δ_2 ↑	δ_3 ↑
In-distribution inference: Training on BV2 training set, test on BV 2 test set								
BITD [8]	AV	0.323	-	2.286	-	0.647	0.834	0.901
AVS-Net	V	0.189	0.233	0.696	0.253	0.745	0.938	**0.978**
AVS-Net	AV	**0.170**	**0.210**	**0.678**	**0.245**	**0.776**	**0.942**	0.976
Zero-shot inference: Training on BV2 training set, test on the **BV1** test set								
AVS-Net	V	0.315	0.518	1.149	0.479	0.414	0.680	0.842
AVS-Net	AV	**0.277**	**0.450**	**1.006**	**0.398**	**0.537**	**0.792**	**0.906**

3.2. Evaluation metrics.

We resize predicted depth maps to the ground truth resolution (720×1280) and only use valid pixels[2]. For evaluation metrics, we adopted the commonly used Abs Rel, Sq Rel, RMSE, RMSE log, and three threshold accuracies $\delta_1 < 1.25$, $\delta_2 < 1.25^2$, $\delta_3 < 1.25^3$.

3.3. Results

We demonstrate enhanced pseudo-dense metric depth using echoes and validate the effectiveness of AVS-Net to retrieve metric factors. We also show contributions of echoes across depth ranges. We ablate only visual scaling AVS-Net (V) wrt. scaling with RGB-echoes AVS-Net (AV).

Enhanced Pseudo-Dense Metric Depth with echoes. In table 1 we show the AVS-Net using echoes surpassed the only visual baseline by a margin of about 10% for both

[1]The rationale behind the introduction of the *pseudo-dense* metric depth is to facilitate differentiation from the final metric depth, i.e. depth map combining metric factors and relative depth

[2]Depth-from-stereo artifacts are zero in the ground truth and filtered, so are pixel larger than the maximum, 12m for BV1 and 30m for BV2.

Table 2. Prediction results on the **BV1 test set**. "w/o" denotes no scaling, *Baseline* denotes only-visual scaling, *Proposed* denotes full Audio-Visual scaling method. Best metrics are bold, best across all models are underlined.

	Models	Scaling Method	**Abs rel** ↓	**sq rel** ↓	**RMSE** ↓	**RMSE(log)** ↓	δ_1 ↑	δ_2 ↑	δ_3 ↑
Relative Depth Estimation	Monodepth2[5]	w/o	0.649	0.931	1.656	1.168	0.008	0.023	0.065
		Baseline	0.312	0.354	0.975	0.498	0.397	0.677	0.827
		Proposed	**0.289**	**0.318**	**0.898**	**0.433**	**0.490**	**0.753**	**0.874**
	MonoVit[12]	w/o	0.568	0.757	1.490	0.998	0.017	0.072	0.283
		Baseline	0.320	0.386	1.002	0.531	0.393	0.673	0.814
		Proposed	**0.297**	**0.355**	**0.931**	**0.465**	**0.484**	**0.746**	**0.860**
	LiteMono[11]	w/o	0.782	1.284	1.933	1.626	0.000	0.000	0.001
		Baseline	0.284	0.313	0.951	0.458	0.436	0.719	0.852
		Proposed	**0.261**	**0.270**	**0.875**	**0.393**	**0.525**	**0.785**	**0.898**
Metric Depth Estimation **Zero-Shot**	ZoeDepth[2]	w/o	1.000	1.630	1.558	0.705	0.107	0.258	0.486
		Baseline	0.285	0.394	1.181	0.455	0.420	0.677	0.829
		Proposed	**0.250**	**0.301**	**1.042**	**0.377**	**0.512**	**0.780**	**0.901**
	Jun et al.,[6]	w/o	0.549	0.597	**1.083**	0.483	0.241	0.532	0.828
		Baseline	0.303	0.437	1.258	0.493	0.387	0.647	0.807
		Proposed	**0.260**	**0.342**	1.129	**0.409**	**0.495**	**0.755**	**0.880**
	NeWCRFs [10]	w/o	1.240	2.485	1.874	0.813	0.094	0.208	0.388
		Baseline	0.309	0.453	1.287	0.505	0.386	0.630	0.793
		Proposed	**0.276**	**0.361**	**1.156**	**0.425**	**0.463**	**0.737**	**0.869**

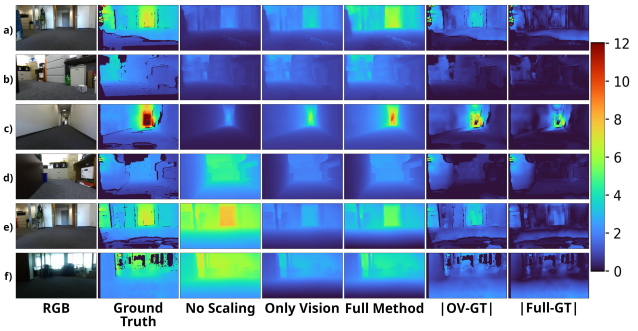

Figure 2. Prediction examples. *Full Method* denotes Audio-Visual scaling. a) to f) are based on Monodepth2, MonoVit, LiteMono, ZoeDepth, NeWCRFs, Jun et al. The last two columns show absolute differences between predictions and ground truth.

Abs Rel and Sq Rel on BV2. Tested zero-shot on BV1, the improvement margins are further enlarged to about 12% for Abs Rel and 13% for Sq Rel. Especially for δ_1 accuracy, using echoes achieved about 23% improvement compared to its only-visual counterpart, demonstrating good generalization. While BV1 and BV2 are published as one dataset, they contain significantly different scenes, depth profiles and sensors used [3]. Interestingly, AVS-Net (V) already significantly outperformed the SOTA method, Parida *et al.*[8] on BV2. This could be attributed to the transfer learning from Midas. Large improvements of AVS-Net (AV) demonstrate efficient use of echoes with this network architecture.

Effectiveness of AVS-Net for metric factors. We evaluate AVS-Net (AV) by improving (or correcting) the scale of several relative depth models [5, 12, 11] trained self-supervised. We also compare zero-shot results when improving pre-trained metric-depth models [2, 6, 10] tested on $BV1$. Because relative depth-trained models are not designed to output metric depth, performance differences are only meaningful between similar methods. For space reasons we still show all methods in tab. 2. Our Audio-Visual scaling method surpassed all the baseline method without echoes, showing good performance providing robust metric scale factors. Compared with visual baselines, improvements of about 10% in Abs Rel and Sq Rel, about 19% for δ_1 are obtained for relative depth models (first 3 models), while for metric-depth models (last 3 models) about 20% improvements for Sq Rel and δ_1 are observed. All relative depth estimation methods were originally trained on KITTI so their original weights were also tested on BV1. Results were inferior and omitted due to space (e.g. Monodepth2 Abs rel (Proposed)=0.300).

Qualitative results in Fig.2 show the visual scaling offers reasonable estimation of scale but is unable to accurately predict distances to obstacles or corridors further away. In contrast, audio-visual scaling resolves distance ambiguity by incorporating information from echoes. This reduces the discrepancy between estimated and actual distances in comparison to the ground truth (last two columns). Please note that the final quality of the metric depth depends also on the relative depth part, thus on different models and training schemes (i.e. Comparing *Full Method* in Fig. 2).

Analysis by depth intervals. We investigate audio-visual performance of all methods for different depth-intervals. For each ground truth depth map $\mathcal{M}_n$, we define

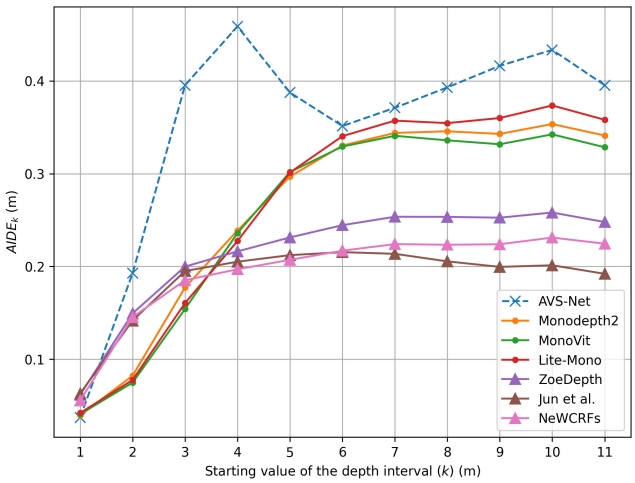

Figure 3. Average improvement using echoes over depth intervals.

the region Ω_n^k as the set of depth values between k and $k+1$:

$$\Omega_n^k = \{(i,j), k < \mathcal{M}_n(i,j) < k+1\}, \quad k \in \mathbb{N} \cap (0, MAX(\mathcal{M}_n)) \tag{1}$$

Then the Average Improvement in Depth Error ($AIDE_k$) can be defined as:

$$\mathcal{E}_n^{AV}(i,j) = \left|\hat{\mathcal{M}}_n^{AV}(i,j) - \mathcal{M}_n(i,j)\right| \tag{2}$$

$$\mathcal{E}_n^{OV}(i,j) = \left|\hat{\mathcal{M}}_n^{OV}(i,j) - \mathcal{M}_n(i,j)\right| \tag{3}$$

$$AIDE_k = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{|\Omega_n^k|}\sum_{(i,j)\in\Omega_n^k}(\mathcal{E}_n^{OV}(i,j) - \mathcal{E}_n^{AV}(i,j)) \tag{4}$$

$\mathcal{E}_n^{AV}$ and $\mathcal{E}_n^{OV}$ denote errors for the Audio-Visual and Only-Visual method respectively, defined as difference between the depth predictions $\hat{\mathcal{M}}_n^{AV}$, $\hat{\mathcal{M}}_n^{OV}$ and the ground truth $\mathcal{M}_n$. n and N denote the sample index and the total number of samples in the test set, respectively.

Figure 3 shows $AIDE_k$ results for each model. Note, the *AVS-Net* produces $\hat{\mathcal{M}}_p$ to scale other methods. Therefore, the metric is calculated with $\hat{\mathcal{M}}_p$ for *AVS-Net* and final scaled depth maps for the others. The results indicate that all methods benefit more from echoes at larger distances. Echoes significantly enhance $\hat{\mathcal{M}}_p$, more than scaling results. This may be due to the coarse-grained scaling by the global median value. Fine-grained scaling methods, such as refinement modules, could further improve performance.

Improvements for self-supervised models (Monodepth2, MonoVit, Lite-Mono) and zero-shot models (ZoeDepth, Jun et al., NeWCRFs) form two distinct clusters. Analysis of estimated depth distributions suggests these differences stem from variations in training methods (disparity vs. depth prediction) and depth distributions of the training datasets.

4. Conclusion

We propose AVS-Net which employs sound for enhanced metric depth estimation. Scale factors are derived from echoes to provide missing scale information for relative depth and rectify pre-trained metric depth models. Experimental results shows the method markedly outperforms the visual-only baseline, showcasing the efficacy of echoes in retrieving scale-correct depth. Scaling the output, the AVS-Net allows for simple integration with other relative depth or metric depth models, thereby making the method applicable in a wide range of scenarios.

Acknowledgement Supported by the ANR Grant No. ANR22-CE94-0003, NSFC Project No. 62432009 and HPC resources from GENCI–IDRIS (Grant 2024-103685)

References

[1] S. F. Bhat, I. Alhashim, and P. Wonka. Localbins: Improving depth estimation by learning local distributions. In *European Conference on Computer Vision*. Springer, 2022. 1

[2] S. F. Bhat, R. Birkl, D. Wofk, P. Wonka, and M. Müller. Zoedepth: Zero-shot transfer by combining relative and metric depth. *arXiv preprint arXiv:2302.12288*, 2023. 1, 2, 3

[3] A. Brunetto, S. Hornauer, X. Y. Stella, and F. Moutarde. The audio-visual batvision dataset for research on sight and sound. In *IEEE/RSJ IROS*. IEEE, 2023. 1, 2, 3

[4] R. Gao, C. Chen, Z. Al-Halah, C. Schissler, and K. Grauman. VisualEchoes: Spatial Image Representation Learning Through Echolocation. In *ECCV 2020*. 1

[5] C. Godard, O. Mac Aodha, M. Firman, and G. J. Brostow. Digging into self-supervised monocular depth estimation. In *IEEE/CVF ICCV*, 2019. 1, 3

[6] J. Jun, J.-H. Lee, C. Lee, and C.-S. Kim. Depth map decomposition for monocular depth estimation. In *IEEE ECCV*. Springer, 2022. 1, 3

[7] X. Liu, A. Brunetto, S. Hornauer, F. Moutarde, and J. Lu. Pano-echo: Panoramic depth prediction enhancement with echo features. In *IEEE CAI*, 2024. 1, 2

[8] K. K. Parida, S. Srivastava, and G. Sharma. Beyond image to depth: Improving depth prediction using echoes. In *IEEE CVPR*, 2021. 1, 2, 3

[9] R. Ranftl, K. Lasinger, D. Hafner, K. Schindler, and V. Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. In *IEEE PAMI*, 2020. 1, 2

[10] W. Yuan, X. Gu, Z. Dai, S. Zhu, and P. Tan. Neural window fully-connected crfs for monocular depth estimation. In *IEEE/CVF CVPR*, 2022. 1, 3

[11] N. Zhang, F. Nex, G. Vosselman, and N. Kerle. Lite-mono: A lightweight cnn and transformer architecture for self-supervised monocular depth estimation. In *IEEE/CVF CVPR*, 2023. 1, 3

[12] C. Zhao, Y. Zhang, M. Poggi, F. Tosi, X. Guo, Z. Zhu, G. Huang, Y. Tang, and S. Mattoccia. Monovit: Self-supervised monocular depth estimation with a vision transformer. In *2022 international conference on 3D vision (3DV)*, pages 668–678. IEEE, 2022. 1, 3