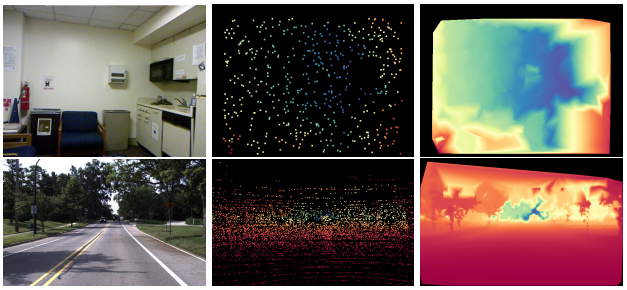
Need for Speed: Zero-Shot Depth Completion with Single-Step Diffusion

Supplementary Material



(a) RGB (b) Sparse (c) Completed

Figure 1. **Barycentric intepolation.** Example of completing sparse depth using Barycentric interpolation within Delaunay triangulation. RGB image is provided here only for reference.

1. Barycentric Interpolation

In the Fig. 1 we provide a visualization of depth completion utilizing barycentric interpolation within Delaunay triangulation. We set values outside of the convex hull to zero. As discussed in our paper, the computationally inexpensive methods like the interpolation produce competitive results for denser depth conditions.

2. Evaluation Under Varying Depth Density

Fig. 2 showcases results under varying depth densities for IBims-1, VOID, NYUv2, ScanNet and DDAD dataset. Evaluation on KITTI dataset was performed only considering all provided LiDAR points.

3. Ablation on Sampling Condition Density

The complete results for sampling condition density during fine-tuning are illustrated in Fig. 3 The evaluation of the models (A), (B) & (C) makes it more apparent that as soon as the models are evaluated on the sparse depth density outside of the training range, the performance drops. Widening the sampling density range does not impact the performance significantly for particular evaluation density. For example, the performance of model (A) is only slightly better compared to Marigold-SSD for the density of 0.16%, the same goes for model (C) compared of Marigold-SSD for density of 0.5%. One exception to this is generally degraded performance of model (C) on VOID dataset. This could be due to less uniform sampling of the depth condition provided in the VOID dataset.

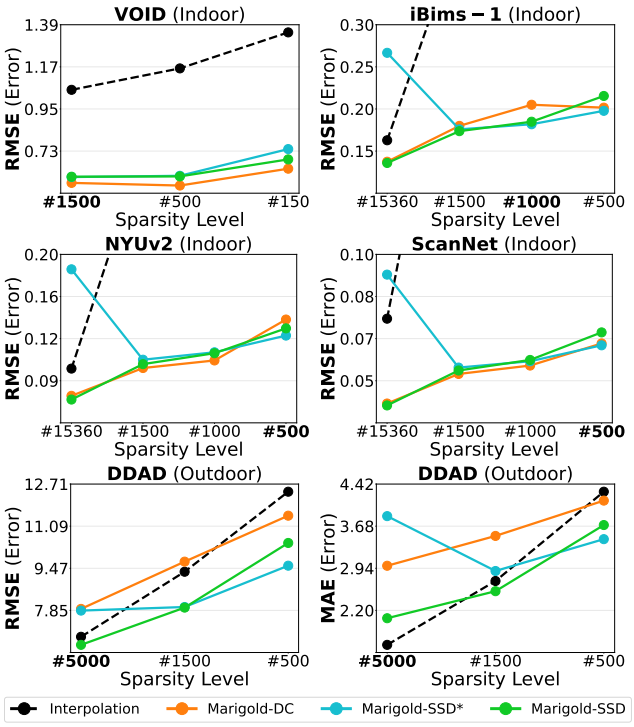


Figure 2. **Evaluation under multiple levels of depth density**, for IBims-1, VOID, NYUv2, ScanNet and DDAD datasets. The density of the depth condition is denoted by number of depth samples (#). Only RMSE is presented for indoor datasets as the trends for MAE are the same. On DDAD dataset at the commonly used sparsity level of 5000 points even sophisticated models can be outperformed by trivial Barycentric interpolation.

Table 1. **Depth boundary evaluation on IBims-1 dataset [2].**

Model	IBims-1 $\mathcal{E}_{\text{DBE}}^{\text{acc}} \downarrow$	IBims-1 $\mathcal{E}_{\text{DBE}}^{\text{comp}} \downarrow$
Marigold-E2E	1.756	10.944
Marigold-DC w/ ensembling	1.537	7.078
Marigold-DC w/o ensembling	1.706	5.524
Marigold-SSD (Ours)	1.768	7.485

4. Accuracy of Depth Boundaries

Additionally, we extend the evaluation protocol by *depth boundary error (DBE)* measures assessing depth boundaries on IBims-1 dataset [2]. The ground-truth for edge evaluation in IBims-1 was established by manually selecting distinct edges from edge hypotheses generated by structured edges [1]. For evaluation, the edges in depth maps Y_{bin} are

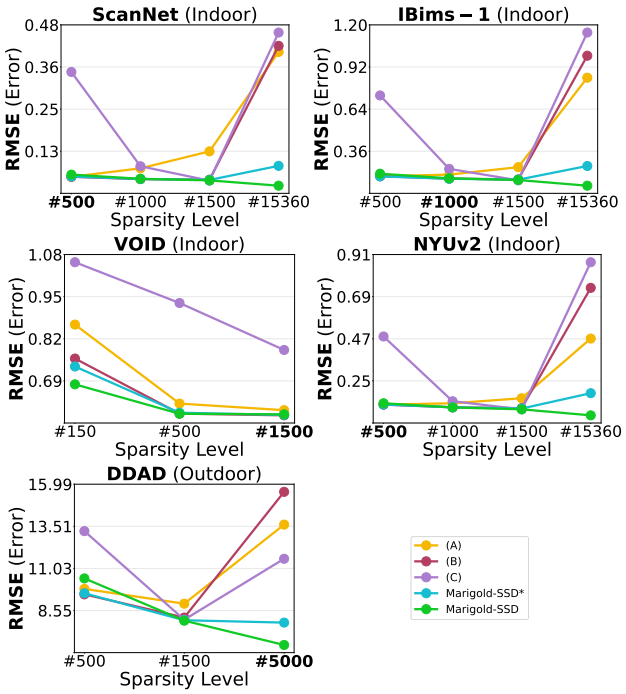


Figure 3. **Sampling Condition Density**. Models (A) & (C) were fine-tuned on constant density levels of 0.16% and 0.5%, model (B) on the narrower density range $[0.16\%, 0.32\%]$ compared to our default settings. Performance degrades when the models evaluated with out-of-domain densities.

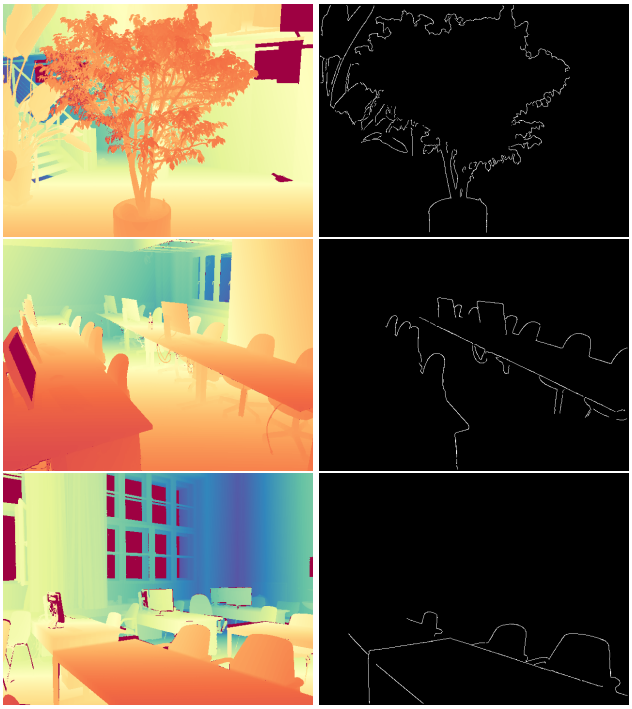
extracted by the same structured edges method and compared with the ground-truth edges Y^*_{bin} via truncated chamfer distance. The accuracy measure:

$$\mathcal{E}^{\text{acc}}_{\text{DBE}}(Y) = \frac{1}{\sum_i \sum_j y_{\text{bin};i,j}} \sum_i \sum_j e^*_{i,j} \cdot y_{\text{bin};i,j} \qquad (1)$$

applies euclidean distance transform $E^* = DT(Y^*_{\text{bin}})$ to the ground-truth edge image ignoring distances larger than 10 pixels. The accuracy measure does not capture missing edges in the predicted depth maps, thus completeness measure is defined as:

$$\mathcal{E}^{\text{comp}}_{\text{DBE}}(Y) = \frac{1}{\sum_i \sum_j y^*_{\text{bin};i,j}} \sum_i \sum_j e_{i,j} \cdot y^*_{\text{bin};i,j} \qquad (2)$$

where the distance image is is computed from the predicted edges $E = DT(Y_{\text{bin}})$. We evaluated our method, Marigold-DC (with and without ensembling) and Marigold-E2E. Results are presented in Tab. 1. Our method achieves accuracy comparable with Marigold-E2E while achieving better completeness. Marigold-DC with ensembling achieves the better accuracy and without ensembling the better completeness. However, we note that the boundary evaluation



(a) Depth (b) Boundaries

Figure 4. **Examples of ground-truth depth maps and boundaries from IBims-1 dataset** for depth and boundary evaluation.

Table 2. In Model (R), conditional decoder is randomly initialized.

Dataset	(R)		**Marigold-SSD**	
	MAE↓	RMSE↓	MAE↓	RMSE↓
ScanNet	0.028	0.070	0.027	0.068
IBims-1	0.065	0.198	0.060	0.185
VOID	0.193	0.611	0.182	0.590
NYUv2	0.054	0.137	0.052	0.134
KITTI	0.432	1.502	0.454	1.496
DDAD	2.196	6.938	2.065	6.522

on IBims-1 dataset is limited to the most significant edges in the scene. This is evident in Fig. 4 which illustrates three examples of the provided depth and boundary ground-truth. Thus, this metric cannot fully capture the fine-grained differences in the detail generation. We provide more samples for the qualitative comparison in Fig. 5 where the differences are more clear especially on plants.

5. Conditional Decoder Weight Initialization

We trained a model where the depth condition encoding layers of our conditional decoder were initialized randomly, instead of initializing weights from VAE encoder when possible. The results are presented in the Tab. 2.

(a) RGB (b) Marigold-SSD (Ours) (c) Marigold-E2E (d) Marigold-DC w/o ens. (e) Marigold-DC w/ ens.

Figure 5. **The methods exhibit different levels of sharpness and detail generation**. All examples are from IBims-1 dataset [2] with a sparsity level of 1000 depth samples.

References

[1] Piotr Dollár and C. Lawrence Zitnick. Fast edge detection using structured forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(8):1558–1570, 2015. 1

[2] Tobias Koch, Lukas Liebel, Friedrich Fraundorfer, and Marco Korner. Evaluation of CNN-based Single-Image Depth Estimation Methods. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, 2018. 1, 3