Embodied Depth Prediction Supplemental Material

Anonymous ICCV submission

Paper ID 11100

In this supplemental document, we provide experimental details of our method (Section A) and additional visualization results (Section B). Please refer to the supplemental webpage for video results.

A. Experimental Details

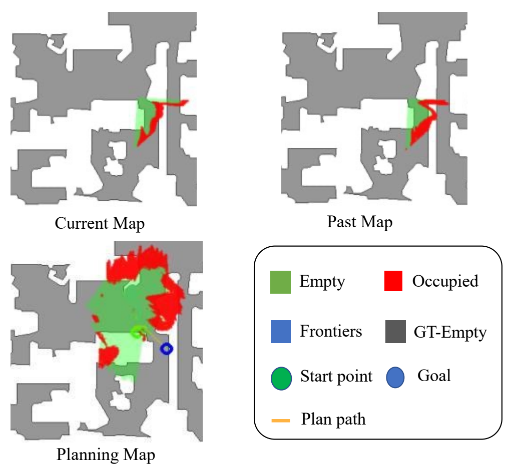
Network Architecture. Our Embodied Depth Network (EDN) takes as input multiple RGB images and their corresponding camera poses and outputs an inverse depth map. To ensure a fair comparison with related networks, we used two past frames to predict the depth. The coarse depth map is obtained as explained in Section 3.5, using a pretrained RAFT [8] model on KITTI [6] dataset. The refinement network architecture is based on a UNet [7], which comprises a ResNet18 [4] encoder to extract features from both the coarse depth map and the current RGB image. These features are then fused using point-wise addition and fed into the decoder, which is a DispNet similar to [10]. The output of the decoder has sigmoid activation layers, while ELU nonlinearities [2] are used elsewhere. We convert the sigmoid output x to depth D using $D = 1/(ax + b)$, where a and b are chosen to ensure that D falls between 0.1 and 20 meters.

Training Details. To train our model, we employ a photometric loss computed from the 4th frame to the current frame for simulations and the 15th frame for real-world data. We set the weight of the smoothness term to 10^{-3}. We use the ADAM optimizer [5] with a learning rate of 10^{-4} and a batch size of 12 on a single Nvidia GTX Titan X. We initialize the model with 6,000 frames of data for warm-up and then train the model every 3,000 frames of data. Once we reach a maximum dataset size of 30,000 frames, we perform an additional 3-epoch training.

B. Additional Results

Active Data Collection. We provide further visualization details on our active data collection strategy, which enables our method to effectively explore and select informative views to improve depth estimation accuracy. The frontier

(a) Illustration of Frontier Exploration. The agent sets the goal to the center of one randomly-picked frontier group which is based on the occupancy map.



(b) Illustration of Depth-Inconsistency Exploration. The agent checks the inconsistent areas between current and past occupancy maps and sets the goal to the center of the depth-inconsistency areas.

Figure 1: **Active Exploration Demo.** We show the top-down maps displaying traversable areas (grey) and untraversable areas (white) to illustrate our method.

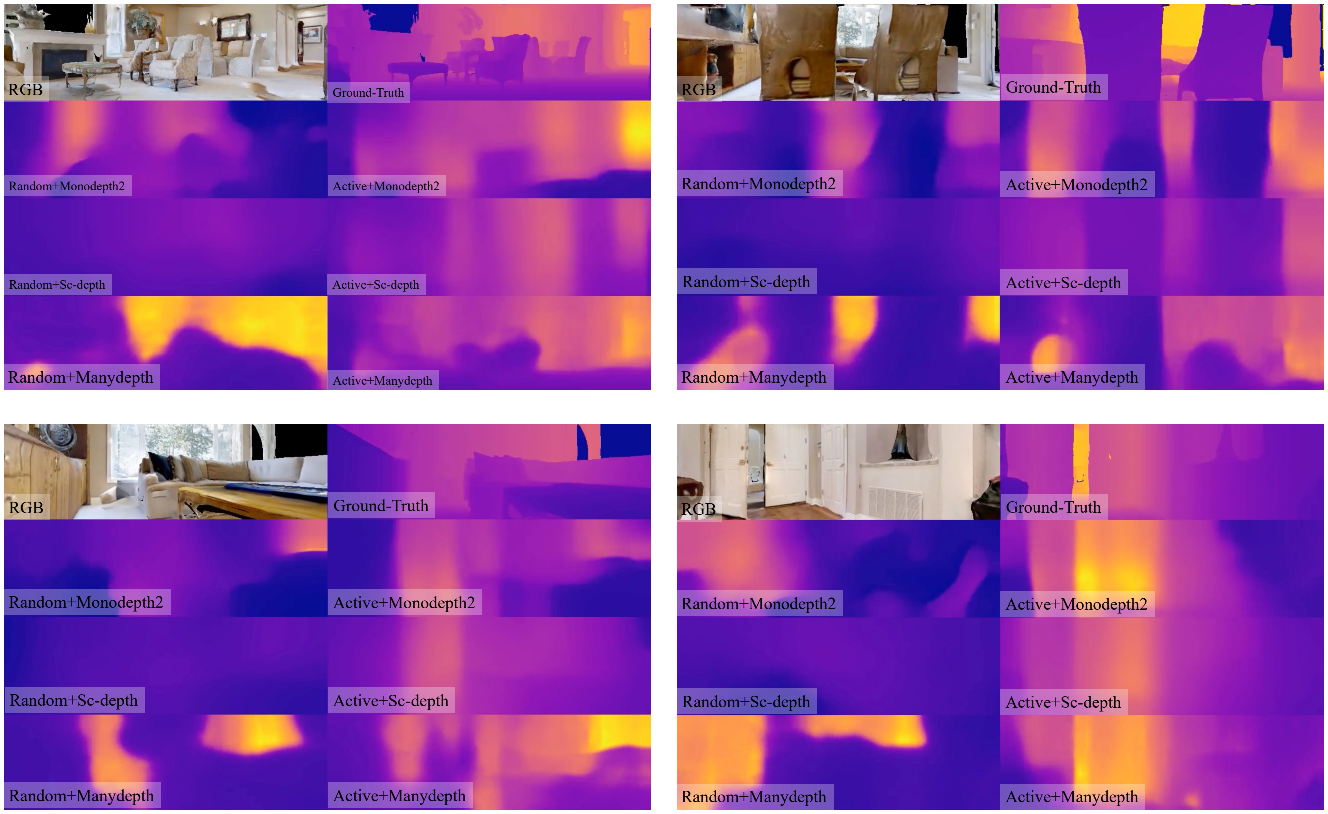


Figure 2: **Depth Predictions in Different Models.** Our active policy improves the performance of all of the baseline models.

exploration expands the distribution of data by selecting views that are close to the current field of view but have not yet been observed. This strategy ensures that the network has access to a diverse set of viewpoints, which can help it learn to generalize better across different scenes. On the other hand, the depth-inconsistency exploration strategy aims to identify areas in the scene where the network is uncertain about its depth predictions. Our active strategy can guide the new data collection even in cases in which the network outputs an ambiguous prediction.

Depth Predictions in Different Models. We compare the performance of different depth estimation models under random and active data collection policies in Fig. 2. Specifically, we visualize the depth predictions of three different models: Monodepth2[3], Sc-Depth[1], and ManyDepth[9]. We find that our active policy consistently improves the performance of all of the baseline models, producing depth predictions that are more accurate than those obtained with random data collection. Our results demonstrate the effectiveness of our proposed active data collection strategy in improving depth estimation accuracy across a range of different models.

References

[1] Jia-Wang Bian, Huangying Zhan, Naiyan Wang, Zhichao Li, Le Zhang, Chunhua Shen, Ming-Ming Cheng, and Ian Reid. Unsupervised scale-consistent depth learning from video. *Int. J. Comput. Vision*, 129(9):2548–2564, sep 2021. 2

[2] Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. Fast and accurate deep network learning by exponential linear units (elus). In *International Conference on Learning Representations (ICLR)*, 2016. 1

[3] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J. Brostow. Digging into self-supervised monocular depth prediction. October 2019. 2

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778. IEEE, June 2016. 1

[5] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations (ICLR)*, 2015. 1

[6] Moritz Menze and Andreas Geiger. Object scene flow for autonomous vehicles. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 1

[7] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, Cham, 2015. Springer International Publishing. 1

[8] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow (extended abstract). In Zhi-Hua Zhou, editor, *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 4839–4843. International Joint Conferences on Artificial Intelligence Organization, 8 2021. Sister Conferences Best Papers. 1

[9] Jamie Watson, Oisin Mac Aodha, Victor Prisacariu, Gabriel Brostow, and Michael Firman. The Temporal Opportunist: Self-Supervised Multi-Frame Monocular Depth. In *Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[10] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G. Lowe. Unsupervised learning of depth and ego-motion from video. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6612–6619, 2017. 1