Supplementary Materials for X-Distill: Improving Self-Supervised Monocular Depth via Cross-Task Distillation

BMVC 2021 Submission # 510

Contents

1 Mathematical Definition of Error Metrics

In this section, we provide the mathematical definitions of the evaluation metrics used in Sec. 4 in the main paper. These are standard metrics used in the literature for evaluating depth estimation performance (e.g., in [1, 2]). For each image, we denote the ground-truth depth map by d, the predicted depth map by $\hat{d}$, the set of valid pixels as Ω where the ground-truth depth values are available, and N_Ω is the number of valid pixels. The errors are then computed as follows for each image:

Absolute Relative Error (Abs Rel): This metric computes the average per-pixel absolute error between the predicted and GT depth values normalized by the GT depth value. The averaging is over the valid pixels in an image:

$$\text{Abs Rel} = \frac{1}{N_\Omega} \sum_{(i,j)\in\Omega} \frac{|d(i,j) - \hat{d}(i,j)|}{d(i,j)}. \tag{1}$$

Squared Relative Error (Sq Rel): This metric computes the average per-pixel squared error between the predicted and GT depth values normalized by the GT depth value. The averaging is over the valid pixels in an image:

$$\text{Sq Rel} = \frac{1}{N_\Omega} \sum_{(i,j)\in\Omega} \frac{(d(i,j) - \hat{d}(i,j))^2}{d(i,j)}. \tag{2}$$

Root Mean Squared Error (RMSE): This metric computes the standard RMSE between the predicted and GT depth values:

$$\text{RMSE} = \sqrt{\frac{\sum_{(i,j)\in\Omega} (d(i,j) - \hat{d}(i,j))^2}{N_\Omega}}. \tag{3}$$

RMSE of Log of Depth Value ($\mathbf{RMSE_{Log}}$): This metric computes the standard RMSE between the log of the predicted depth values and the log of the GT depth values:

$$\text{RMSE}_{\text{Log}} = \sqrt{\frac{\sum_{(i,j)\in\Omega} (\log(d(i,j)) - \log(\hat{d}(i,j)))^2}{N_\Omega}}. \tag{4}$$

Classification Metrics (δ_1, δ_2, δ_3): This set of metrics quantify the closeness between the predicted and GT depth values by measuring their ratio. It computes the percentage of pixels whose ratios are within a certain range from 1. More specifically, we first define:

$$\Omega_k = \{(i,j) \,|\, \max\{\frac{d(i,j)}{\hat{d}(i,j)}, \frac{\hat{d}(i,j)}{d(i,j)}\} \leq 1.25^k \text{ and } (i,j) \in \Omega\}, \tag{5}$$

which is the set of valid pixels whose prediction-GT and GT-prediction ratios are within 1.25^k. The metrics are then given as follows, with $\mathbf{1}_{\Omega_k}$ being the indicator function:

$$\delta_k = \frac{\sum_{(i,j)\in\Omega} \mathbf{1}_{\Omega_k}(i,j)}{N_\Omega}, \ \forall k \in \{1,2,3\}. \tag{6}$$

2 Additional Visual Comparisons with SOTA

In this section, we provide additional visualization examples based on KITTI test data, and compare with the state-of-the-art (SOTA) methods of Monodepth2 [1] and HR-Depth [2].

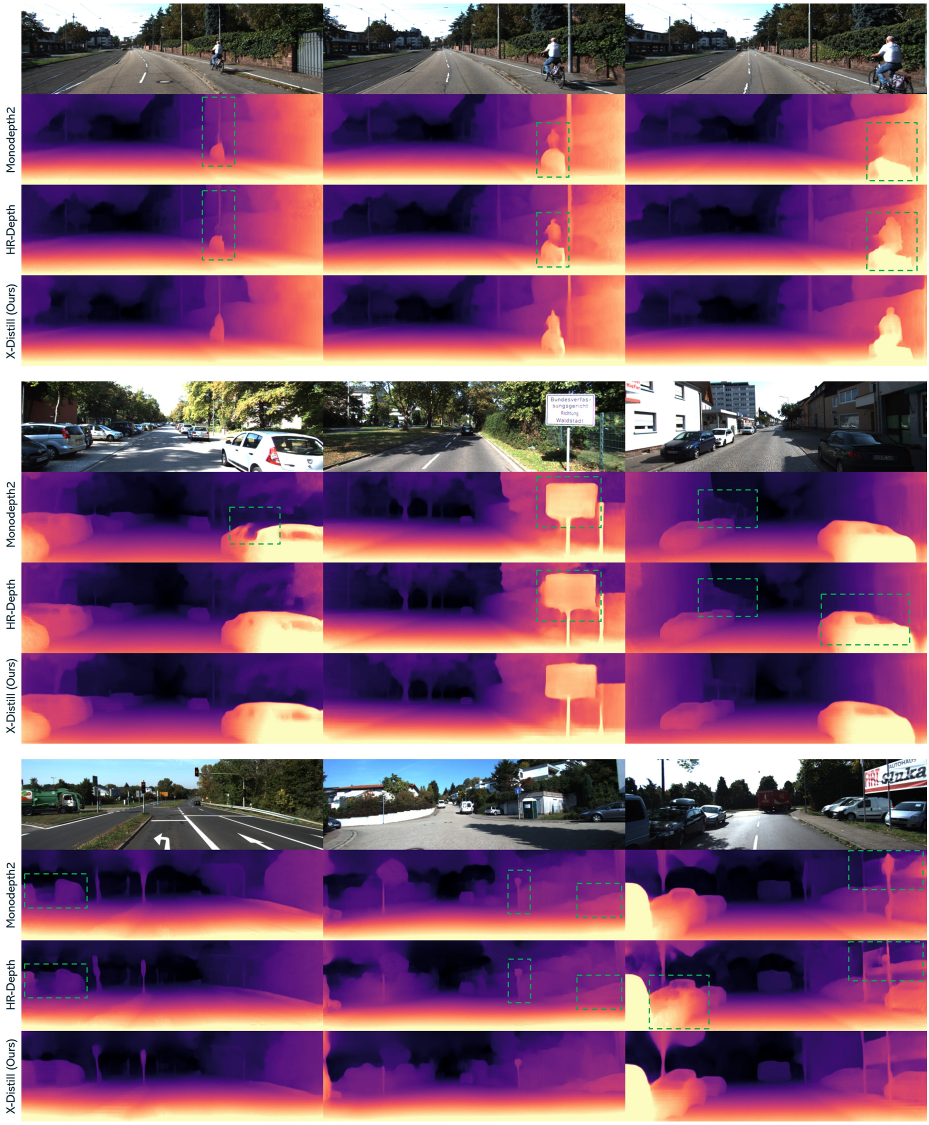

Figure 1: Sample depth estimation results. In each comparison, the second, third, and fourth rows show the estimated depth maps generated by Monodepth2 [1], HR-Depth [2], and our proposed X-Distill approach, respectively. The green boxes indicate sample regions where our method considerably improves the estimation quality.

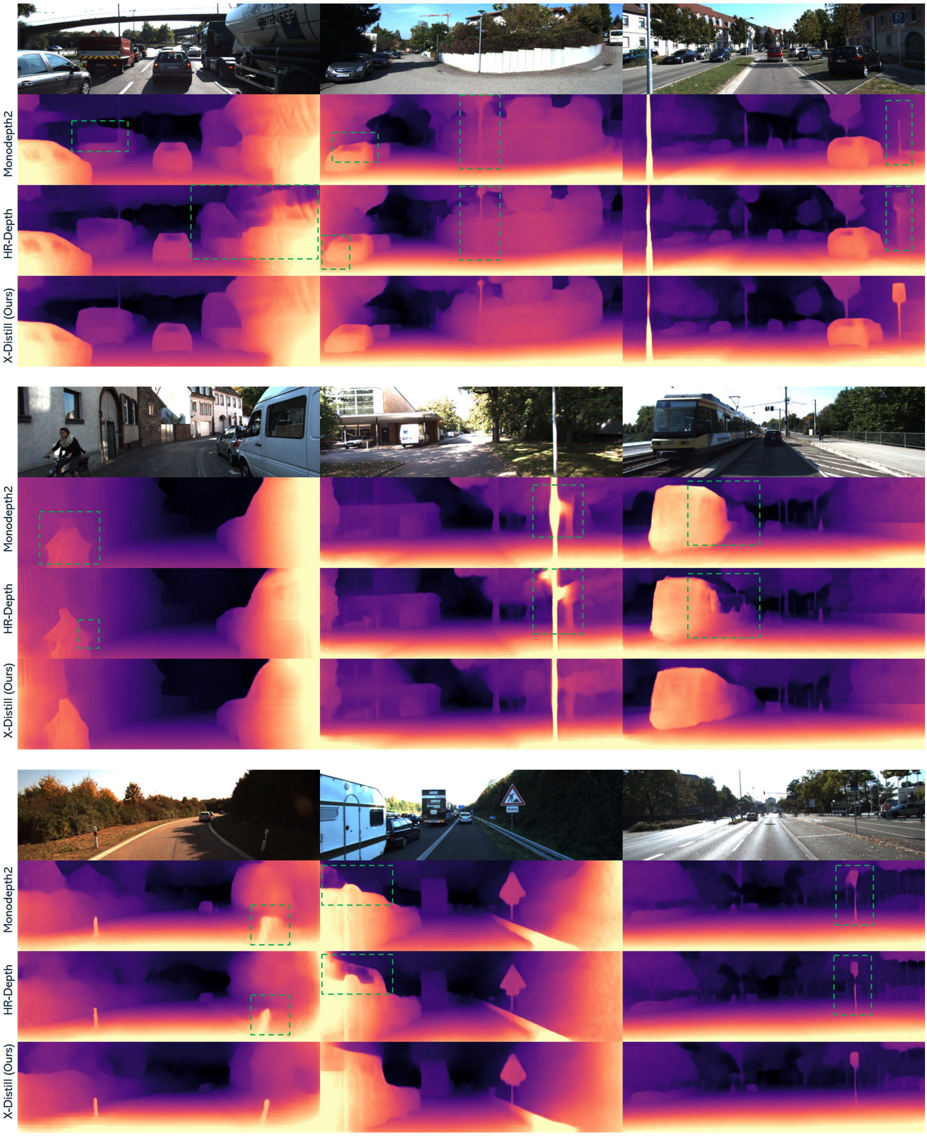

Figure 2: Sample depth estimation results. In each comparison, the second, third, and fourth rows show the estimated depth maps generated by Monodepth2 [1], HR-Depth [2], and our proposed X-Distill approach, respectively. The green boxes indicate sample regions where our method considerably improves the estimation quality.

References

[1] C. Godard, O. Mac Aodha, M. Firman, and G. J. Brostow. Digging into self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019.

[2] X. Lyu, L. Liu, M. Wang, X. Kong, L.a Liu, Y. Liu, X. Chen, and Y. Yuan. HR-Depth:

High resolution self-supervised monocular depth estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021.