Supplementary Materials for
PolyMaX: General Dense Prediction with Mask Transformer

Supplementary Materials

In the supplementary materials, we provide additional information as listed below:

- Sec. A provides detailed training protocol used in the experiments.
- Sec. B provides additional ablations studies.
- Sec. C provides more visualizations of (1) model predictions, (2) failure modes, (3) learned probability distribution maps, and (4) our generated high-quality pseudo-labels for Taskonomy semantic segmentation.

A Training Protocol

The training configurations of PolyMaX closely follow kMaX-DeepLab, including the regularization, drop path [3], color jitting [2], AdamW optimizer [4, 7] with weight decay 0.05, and learning rate multiplier 0.1 for backbone. Additionally, for depth estimation and surface normal, we follow the data preprocessing in [6], except that we disable random scaling and rotation for surface normal.

B Additional Ablation Studies

Impact of Cluster Granularity We analyze the impact of cluster granularity (*i.e.*, K cluster centers) for depth estimation and surface normal, which are presented in Tab. 1 and Tab. 2. Note that, we skip this analysis for semantic segmentation, as we can simply assign the number of clusters as the number of classes. In both Tab. 1 and Tab. 2, we observe that the cluster granularity does not have a significant impact on the model performance on either benchmarks. Among the different cluster settings, 16 clusters and 8 clusters perform the best for depth estimation and for surface normal, respectively.

K	RMS $\downarrow$	A.Rel $\downarrow$	Log$_{10}$ $\downarrow$	$\delta_1 \uparrow$	$\delta_2 \uparrow$	$\delta_3 \uparrow$
4	0.2544	0.0689	0.0295	96.65	99.53	**99.91**
8	0.2578	0.0691	0.0296	96.64	99.58	99.89
16	**0.2499**	**0.0670**	**0.0288**	**96.90**	99.58	99.90
32	0.2520	0.0685	0.0293	96.44	99.56	**99.91**
64	0.2537	0.0688	0.0295	96.77	**99.61**	99.90

Table 1. **Impact of number of clusters (K) on depth estimation.**

K	Mean $\downarrow$	Med $\downarrow$	RMS $\downarrow$	$\delta_1 \uparrow$	$\delta_2 \uparrow$	$\delta_3 \uparrow$
4	13.10	**7.075**	0.2046	65.74	82.19	87.74
8	**13.09**	7.117	**0.2040**	65.66	**82.28**	**87.83**
16	13.15	7.111	0.2051	65.70	82.17	87.73
32	13.11	**7.075**	0.2048	**65.75**	82.23	87.77

Table 2. **Impact of number of clusters (K) on surface normal.**

C Additional Visualization

Model Predictions In Fig. 1, we show more model predictions of semantic segmentation, depth estimation, and surface normal prediction. As shown in the figure, our proposed PolyMaX can capture fine details on scenes with complex structures.

Failure Modes To better understand the limitations of the proposed model, we also look into the failure modes. As shown in Fig. 2, PolyMaX struggles to predict the depth and surface normal for transparent and reflective objects, which are the most challenging issues in the tasks of depth and surface normal estimation. The difficulties can also be reflected by the unreliable ground-truth annotations for those cases. In Fig. 3, our model sometimes predicts over-smoothed depth and surface normal results. The findings of [1, 8] (*e.g.*, a better loss function) may alleviate this issue, which is left for future exploration.

Probability Distribution Maps We provide additional visualizations of the learned probability distribution maps for depth estimation and surface normal prediction in Fig. 4 and Fig. 5, respectively. As shown in the figures, the learned probability distribution maps effectively cluster pixels for different distances (for depth task) or angles (for surface normal task).

Taskonomy Pseudo-Labels In Fig. 6, we show additional visualization of the generated high-quality pseudo-labels for Taskonomy semantic segmentation.

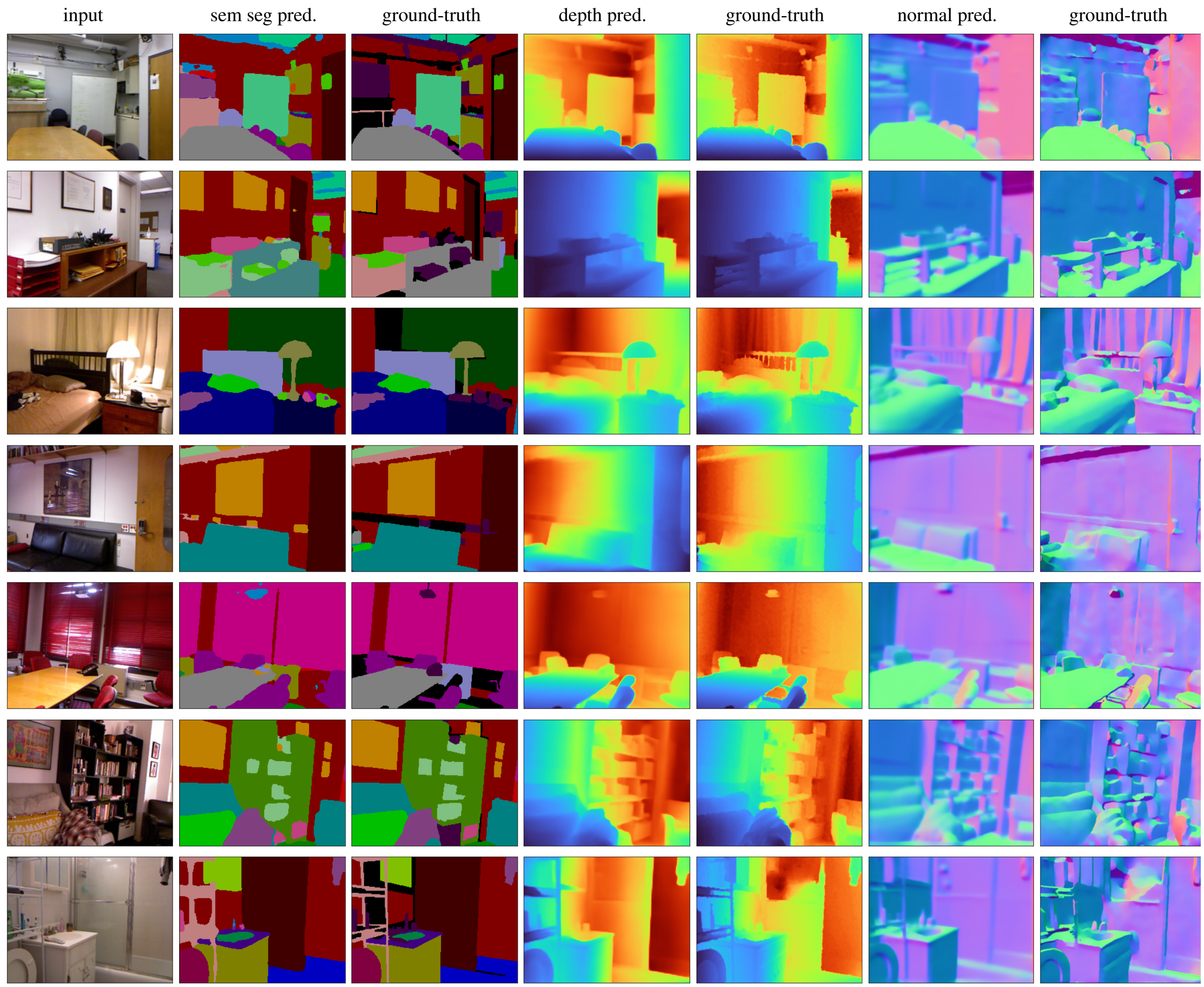

Figure 1. **Visualization of model inputs and outputs for semantic segmentation, depth estimation and normal prediction.** PolyMaX is capable of capturing fine details on scenes with complex structures. Interestingly, as shown in the bottom row, PolyMaX can even reasonably estimate the depth for the glass door, where depth models typically struggle.

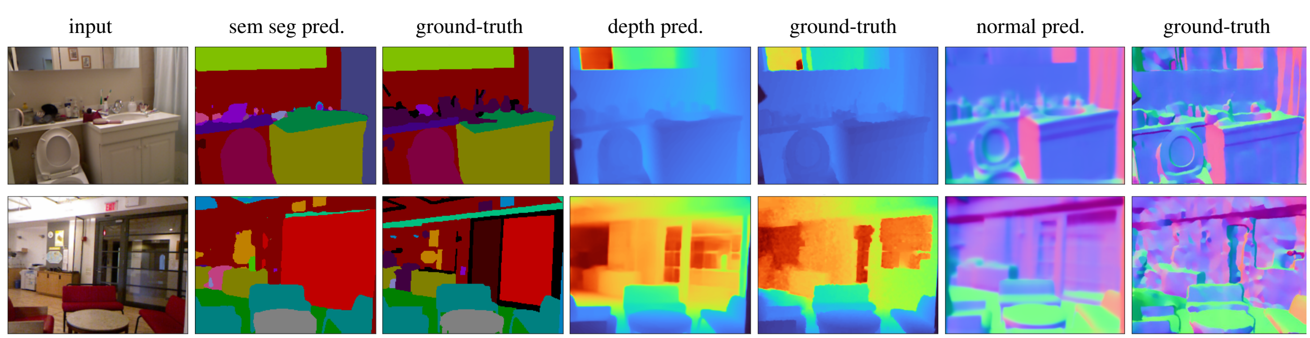

Figure 2. [**Failure mode**] PolyMaX still has difficulties with correctly predicting the depth and surface normal for transparent and reflective surfaces (e.g. mirror in first row, glass in second row). These are well-known challenges for such tasks, especially the ground-truths in these scenarios are also often unreliable, as shown in these examples.

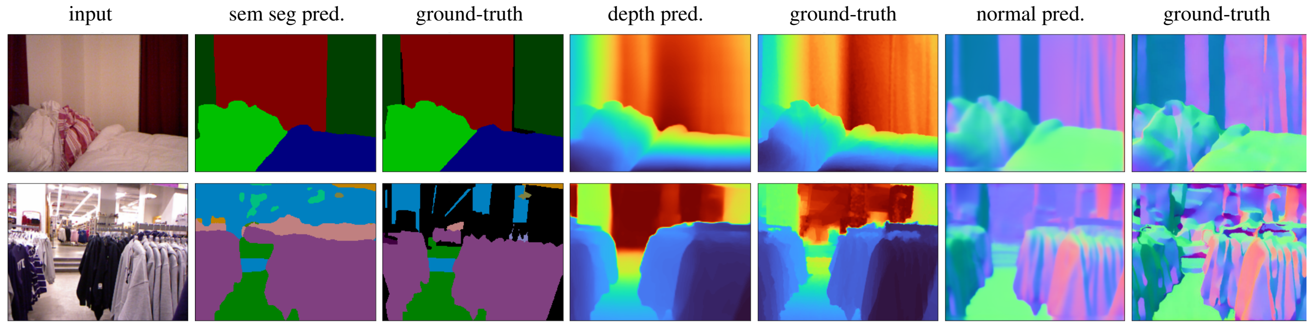

Figure 3. [**Failure mode**] Although PolyMaX achieves superior performance on all three benchmarks on NYUD-v2 dataset, we observe that it still suffers from the over-smoothness issue for depth estimation and surface normal tasks, which other prior works [1, 8] attempt to tackle.

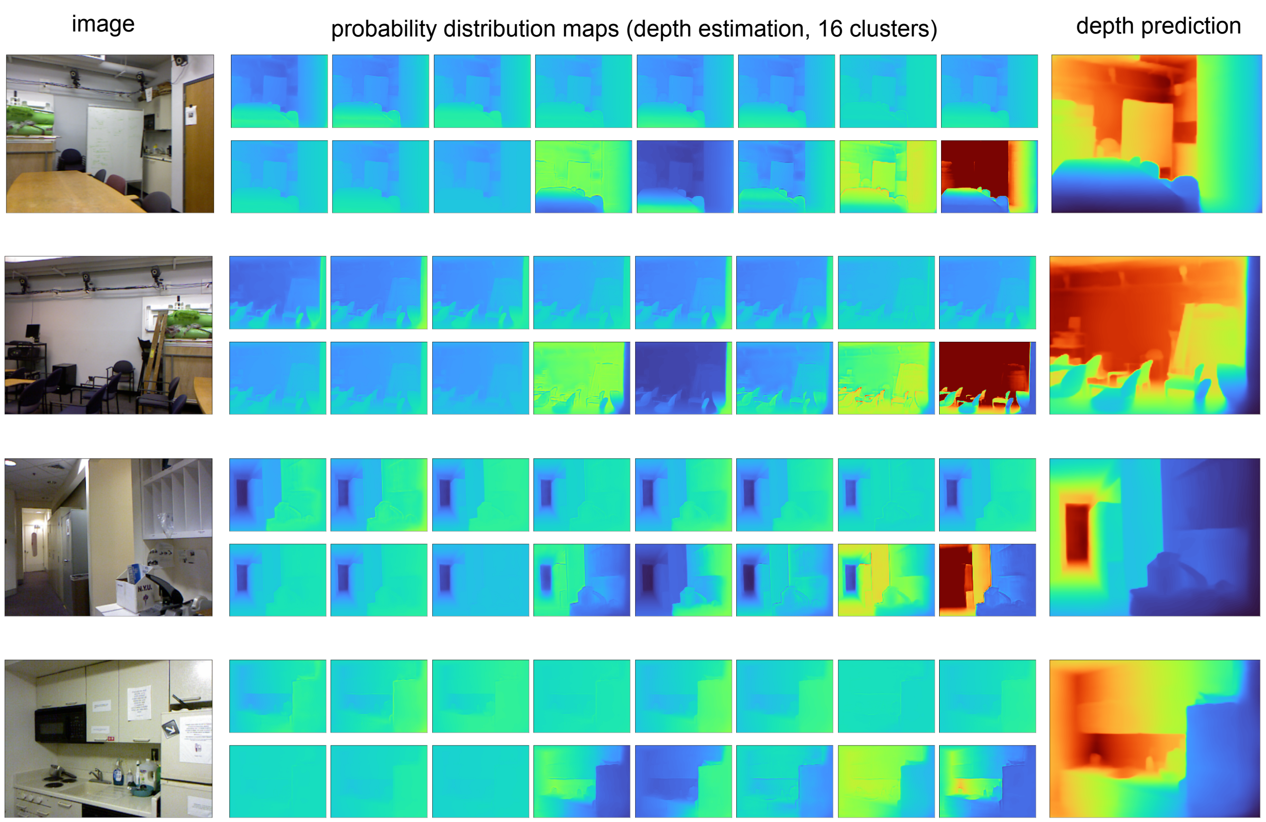

Figure 4. **Additional visualization of probability distribution maps for depth estimation**. Despite of the redundancy in the 16 probability distribution maps, the unique ones clearly demonstrate that the pixels are clustered as closest, mid-range, and furthest distances, which validate the effectiveness of PolyMaX with cluster-prediction paradigm.

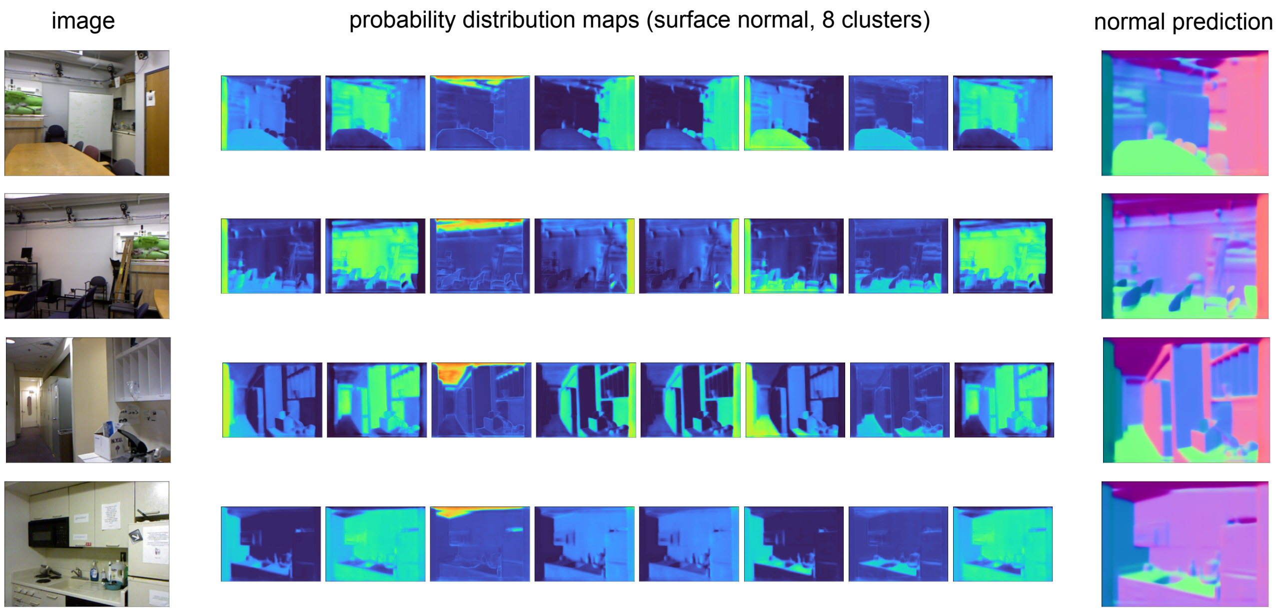

Figure 5. **Additional visualization of probability distribution maps for surface normal prediction**. These probability maps highlight regions with different angles, demonstrating PolyMaX is capable of clustering pixels based on the normal directions.

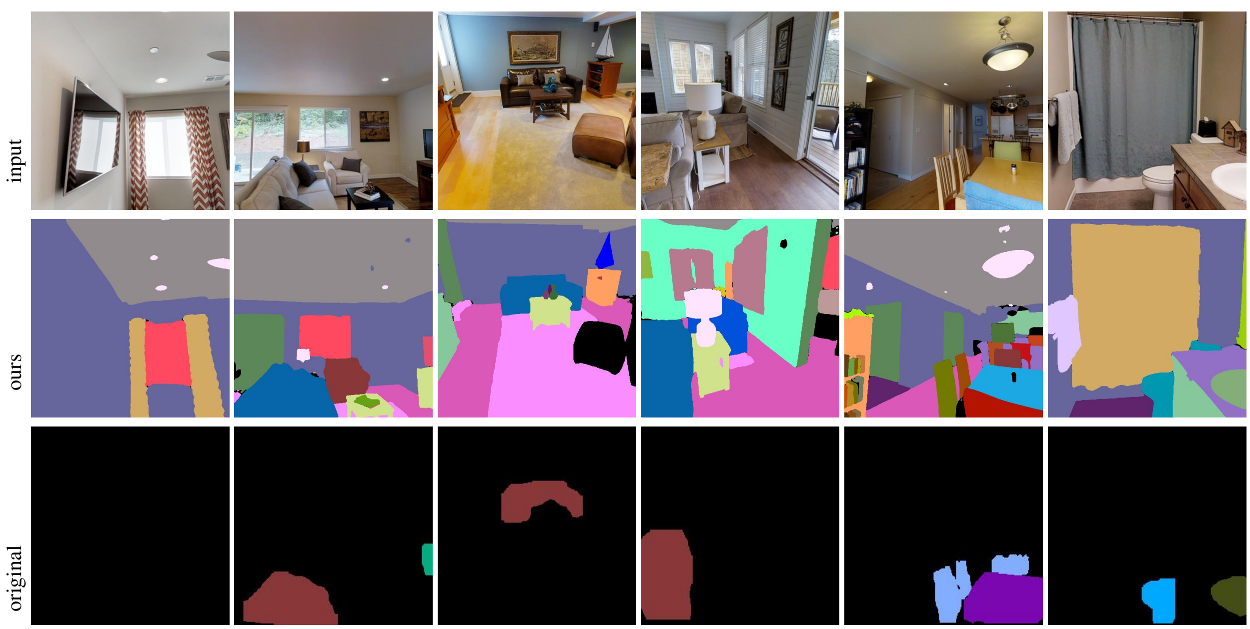

Figure 6. **Additional visualization of Taskonomy pseudo-labels: ours (middle) *vs*. original ones by Li *et al*. [5] (bottom).** Our pseudo-labels demonstrate higher quality than the existing ones.

References

[1] Gwangbin Bae, Ignas Budvytis, and Roberto Cipolla. Estimating and exploiting the aleatoric uncertainty in surface normal estimation. In *ICCV*, 2021. 1, 3

[2] Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation policies from data. In *CVPR*, 2019. 1

[3] Gao Huang, Yu Sun, Zhuang Liu, Daniel Sedra, and Kilian Q Weinberger. Deep networks with stochastic depth. In *ECCV*, 2016. 1

[4] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 1

[5] Yi Li, Haozhi Qi, Jifeng Dai, Xiangyang Ji, and Yichen Wei. Fully convolutional instance-aware semantic segmentation. In *CVPR*, 2017. 4

[6] Zhenyu Li, Xuyang Wang, Xianming Liu, and Junjun Jiang. Binsformer: Revisiting adaptive bins for monocular depth estimation. *arXiv:2204.00987*, 2022. 1

[7] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019. 1

[8] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE TPAMI*, 44(3), 2022. 1, 3