AdaBins: Depth Estimation Using Adaptive Bins (Supplementary Material)

A. Geometric consistency

We provide a qualitative evaluation of the geometric consistency of depth maps predicted by our model. Surface normal maps provide a good way to visualize the orientation and texture details of surfaces present in the scene. Fig F.1 shows the visualization of the normals extracted from the depth maps for our model and for DAV [2] and BTS [3]. Although the orientations predicted by DAV seems to be consistent, the texture details are almost completely lost. BTS, on the other hand, preserves the texture but sometimes results in erroneous orientation details. Our method exhibits detailed texture and consistent orientations without explicitly imposing geometric constraints, such as co-planarity, used by other methods [2, 3].

B. Generlization analysis

Here we qualitatively analyze the capability of our method to generalise to unseen data. We use the models (AdaBins and BTS [3]) trained on NYU-Depth-v2 [4] but show predictions on SUN RGB-D [5] dataset in Fig F.2. Depth maps predicted by BTS have conspicuous artifacts whereas our method provides consistent results on the unseen data.

C. More results on KITTI

Fig F.3 shows a qualitative comparison of BTS [3] and our method on the KITTI dataset. For better visualization, we have removed the sky regions from the visualized depth maps using segmentation masks predicted by a pretrained segmentation model [1]. We can observe that our method demonstrates superior performance particularly in predicting extents and edges of the on-road vehicles, sign-boards and thin poles. Additionally, BTS tends to blend the farther away objects with the background whereas our method preserves the structure with clear separation.

D. MLP head details

We use a three-layer MLP on the first output embedding of the transformer in the mini-ViT module. The architecture details and the parameters used are given in Table D.1.

Layer	Input dimension	Output dimension	Activation
FC	E	256	LeakyReLU (negative_slope=0.01)
FC	256	256	LeakyReLU (negative_slope=0.01)
FC	256	N	-

Table D.1: Architecture details of MLP head. FC: Fully Connected layer, E: Embedding dimension, N: Number of bins.

Variant	$\delta_1\uparrow$	$\delta_2\uparrow$	$\delta_3\uparrow$	REL $\downarrow$	RMS $\downarrow$
Base + R	0.881	0.980	0.996	0.111	0.419
Base + Uniform-Fix-HR	0.892	0.981	0.995	0.107	0.383
Base + Log-Fix-HR	0.896	0.981	0.995	0.108	0.379
Base + Train-Fix-HR	0.893	0.981	0.995	0.109	0.381
Base + AdaBins-HR	**0.903**	**0.984**	**0.997**	**0.103**	**0.364**
1. mViT @ bottleneck	0.885	0.980	**0.996**	0.110	0.416
2. $-$mViT + GPool	**0.896**	**0.983**	**0.996**	**0.107**	**0.370**
3. mViT only ($-$ adaptive bins)	0.892	0.982	0.995	0.108	0.386

Table E.1: Comparison of different variants.

E. Additional ablation

In order to further demonstrate the importance of various components used in our final design, we design other variants of the AdaBins architecture and study their performance. These variants are listed as follows:

1. **mViT at the bottleneck.** In Section 3.3 of the main text, we postulated that global attention at high-resolution is the key to effectively predict the adaptive bin centers. We verify this by moving the mViT block to the bottleneck of the encoder-decoder architecture and study the performance. Note that here we set patch size, $P = 1$.

2. **Adaptive bins via global pooling.** In this design, we study the contribution of mViT module. Specifically, we remove the mViT block and predict the adaptive bin centers using Global Pooling on the decoder features followed by an MLP.

3. **mViT only.** Here, we use the encoder-decoder archi-

tecture followed by mViT without the adaptive bins head. This architecture is equivalent to "Base + mViT + Uniform-Fix-HR" according to nomenclature defined in Table 6 in the main paper.

Performances of the above listed variants trained on NYU-Depth-v2 dataset are listed in Table E.1. We repeat Table 6 of the main paper here for convenience. We can observe that using the mViT module at the bottleneck deteriorates the performance. Furthermore, it can be observed from Table E.1 that the order of importance of the components in our final AdaBins architecture is :

$$\text{global processing} > \text{adaptive bins} > \text{ViT}$$

This validates our hypothesis that global attention at high-resolution is an important factor and adaptive bins is the main component leading to the state-of-the-art performance of our AdaBins design.

References

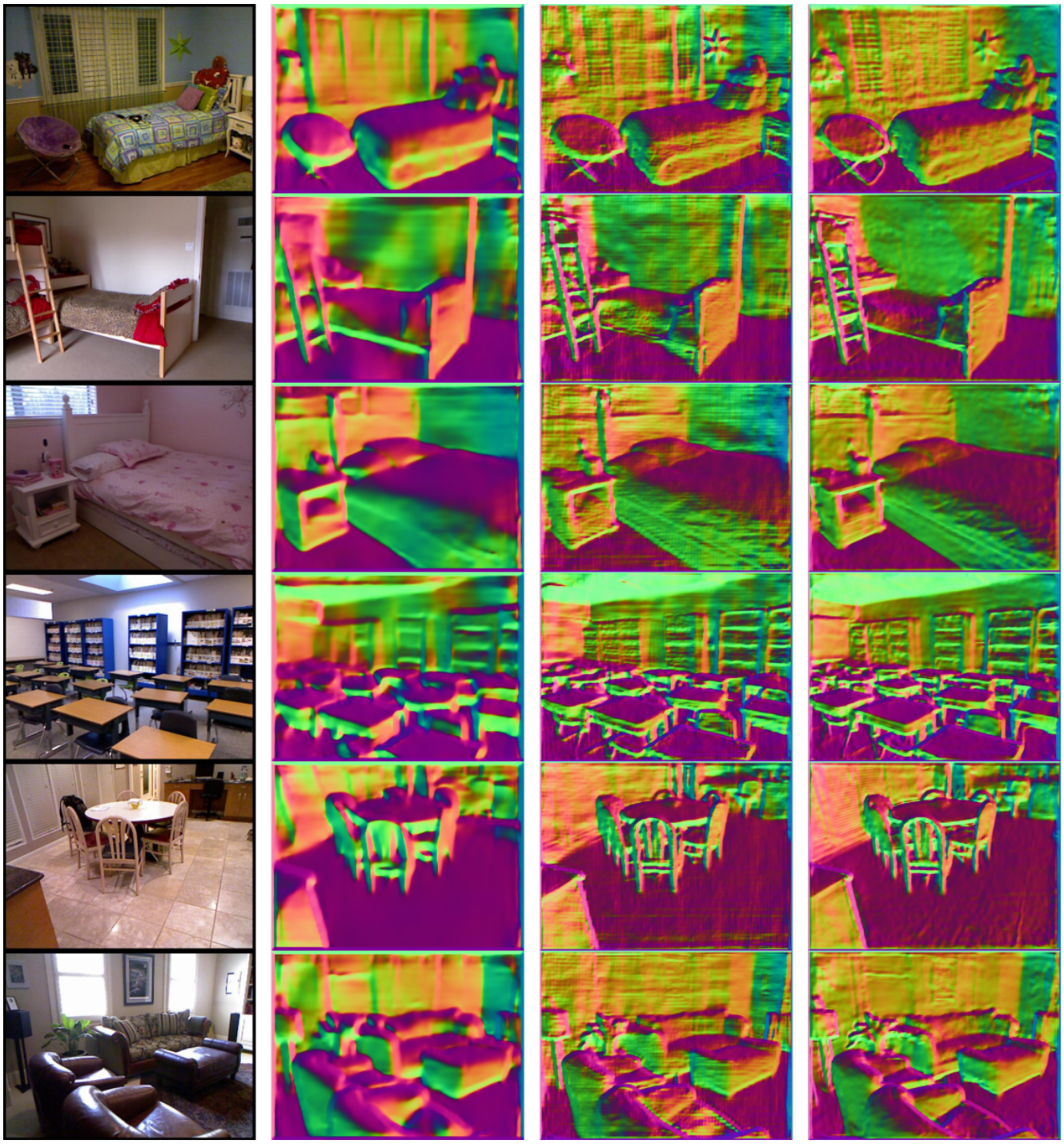
[1] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*, 2018. 1

[2] Lam Huynh, Phong Nguyen-Ha, Jiri Matas, Esa Rahtu, and Janne Heikkila. Guiding monocular depth estimation using depth-attention volume. *arXiv preprint arXiv:2004.02760*, 2020. 1, 3

[3] Jin Han Lee, Myung-Kyu Han, Dong Wook Ko, and Il Hong Suh. From big to small: Multi-scale local planar guidance for monocular depth estimation. *arXiv preprint arXiv:1907.10326*, 2019. 1, 3, 4, 5

[4] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *Computer Vision – ECCV 2012*, pages 746–760, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg. 1

[5] S. Song, S. P. Lichtenberg, and J. Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 567–576, 2015. 1



RGB DAV [2] BTS [3] Ours

Figure F.1: Visualization of surface normals extracted from predicted depth maps.

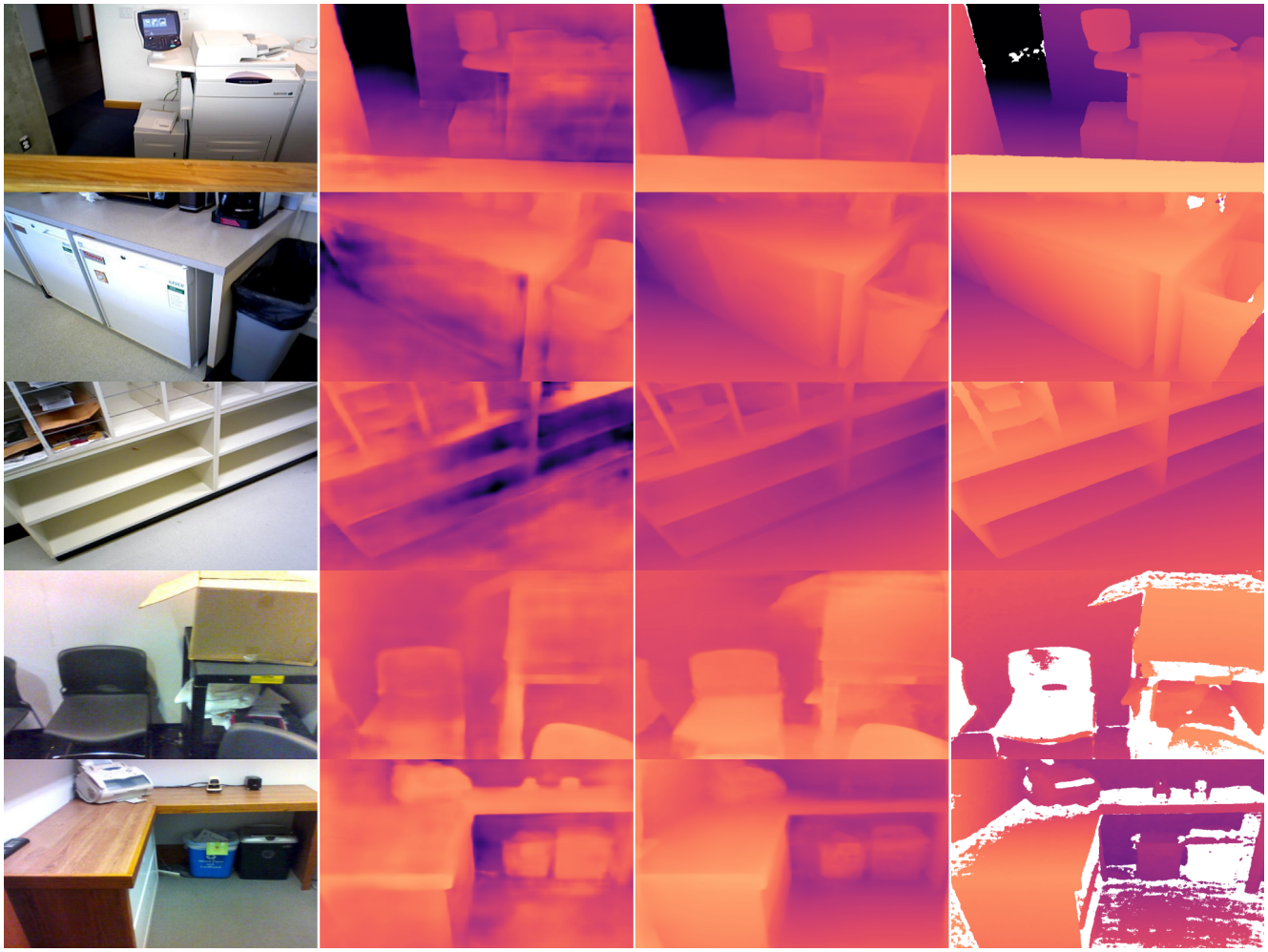
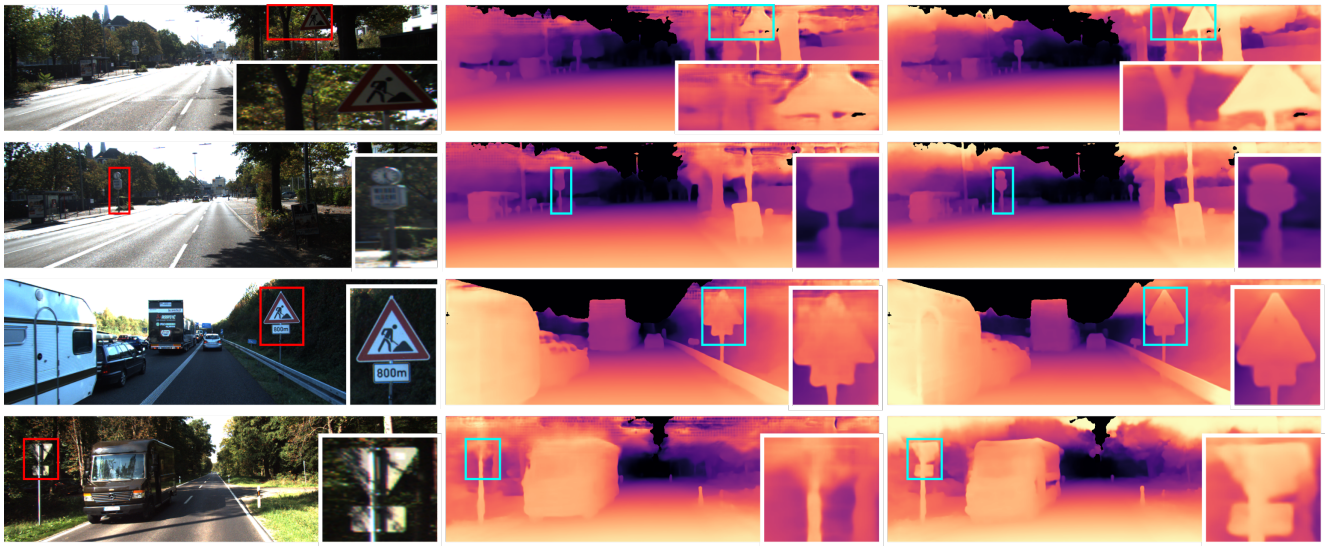


Figure F.2: Qualitative comparison of generalization from NYU-Depth-v2 to SUN RGB-D dataset. Darker pixels are farther. Missing ground truth values are shown in white.



RGB BTS [3] Ours

Figure F.3: Qualitative comparison on KITTI dataset.