Vision Transformers for Dense Prediction – Supplemental Material

1. Architecture details

We provide additional technical details in this section.

Hybrid encoder. The hybrid encoder is based on a pre-activation ResNet50 with group norm and weight standardization [3]. It defines four stages after the initial stem, each of which downsamples the representation before applying multiple ResNet blocks. We refer by RN to the output of the N-th stage. DPT-Hybrid thus taps skip connections after the first ($R0$) and second stage ($R1$).

Residual convolutional units. Figure A1 (a) shows a schematic overview of the residual convolutional units [4] that are used in the decoder. Batch normalization is used for semantic segmentation but is disabled for monocular depth estimation. When using batch normalization, we disable biases in the preceding convolutional layer.

Monocular depth estimation head. The output head for monocular depth estimation is shown in Figure A1 (b). The initial convolution halves the feature dimensions, while the second convolution has an output dimension of 32. The final linear layer projects this representation to a non-negative scalar that represent the inverse depth prediction for every pixel. Bilinear interpolation is used to upsample the representation.

Semantic segmentation head. The output head for semantic segmentation is shown in Figure A1 (c). the first convolutional block preserves the feature dimension, while the final linear layer projects the representation to the number of output classes. Dropout is used with a rate of 0.1. We use bilinear interpolation for the final upsampling operation. The prediction thus represents the per-pixel logits of the classes.

2. Additional results

We provide additional qualitative and quantitative results in this section.

Monocular depth estimation. We notice that the biggest gains in performance for zero-shot transfer were achieved for datasets that feature dense, high-resolution evaluations [1, 2, 5]. This could be explained by more fine-grained predictions. Visual inspection of sample results (see Figure A3) from these datasets confirms this intuition. We observe more details and also better global depth arrangement in DPT predictions when compared to the fully-convolutional baseline. Note that results for DPT and MiDaS are computed at the same input resolution (384 pixels).

	Train. set	DIW	ETH3D	Sintel	KITTI	NYU	TUM
DPT-Large	MIX 5	10.88	0.107	0.309	10.26	10.66	14.31
DPT-Hybrid	MIX 5	11.26	0.104	0.295	17.74	11.43	15.08
MiDaS	MIX 5	12.46	0.129	0.327	23.90	9.55	24.29

Table A1. Additional results on monocular depth estimation.

Additional quantitative results for general-purpose monocular depth prediction when training on the smaller MIX 5 dataset are shown in Table A1.

Semantic segmentation. We show per-class IoU scores for the ADE20K validation set in Figure A2. While we observe a general trend of an improvement in per-class IoU in comparison to the baseline [6], we do not observe a strong pattern across classes.

Attention maps. We show attention maps from different encoder layers in Figures A4 and A5. In both cases, we show results from the monocular depth estimation models. We visualize the attention of two reference tokens (upper left corner and lower right corner, respectively) to all other tokens in the image across various layers in the encoder. We show the average attention over all 12 attention heads.

We observe the tendency that attention is spatially more localized close to the reference token in shallow layers (leftmost columns), whereas deeper layers (rightmost columns) frequently attend across the whole image.

References

[1] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black. A naturalistic open source movie for optical flow evaluation. In *ECCV*, 2012.

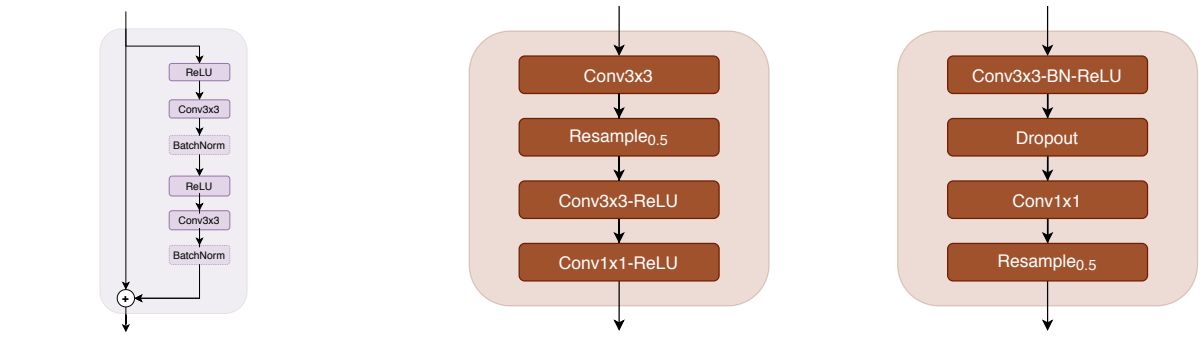
[2] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? The KITTI vision benchmark suite. In *CVPR*, 2012.

[3] Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby. Big transfer (bit): General visual representation learning. In *ECCV*, 2020.

[4] Guosheng Lin, Anton Milan, Chunhua Shen, and Ian D. Reid. RefineNet: Multi-path refinement networks for high-resolution semantic segmentation. In *CVPR*, 2017.

[5] Thomas Schöps, Johannes L. Schönberger, Silvano Galliani, Torsten Sattler, Konrad Schindler, Marc Pollefeys, and Andreas Geiger. A multi-view stereo benchmark with high-resolution images and multi-camera videos. In *CVPR*, 2017.

[6] Hang Zhang, Chongruo Wu, Zhongyue Zhang, Yi Zhu, Zhi Zhang, Haibin Lin, Yue Sun, Tong He, Jonas Muller, R. Manmatha, Mu Li, and Alexander Smola. ResNeSt: Split-attention networks. *arXiv preprint arXiv:2004.08955*, 2020.



(a) Residual Convolutional Unit [4] (b) Monocular depth estimation head (c) Semantic segmentation head

Figure A1. Schematics of different architecture blocks.

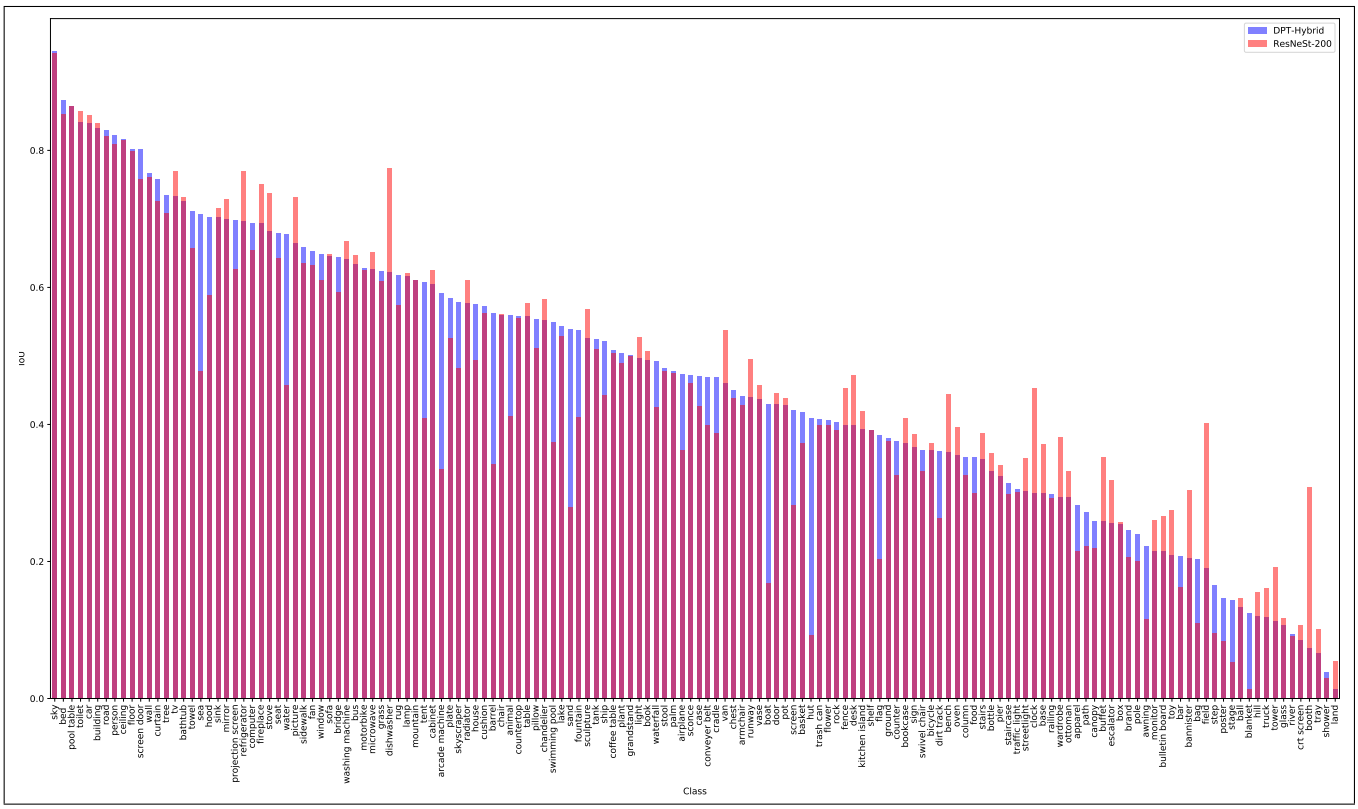


Figure A2. Per class IoU on ADE20K.

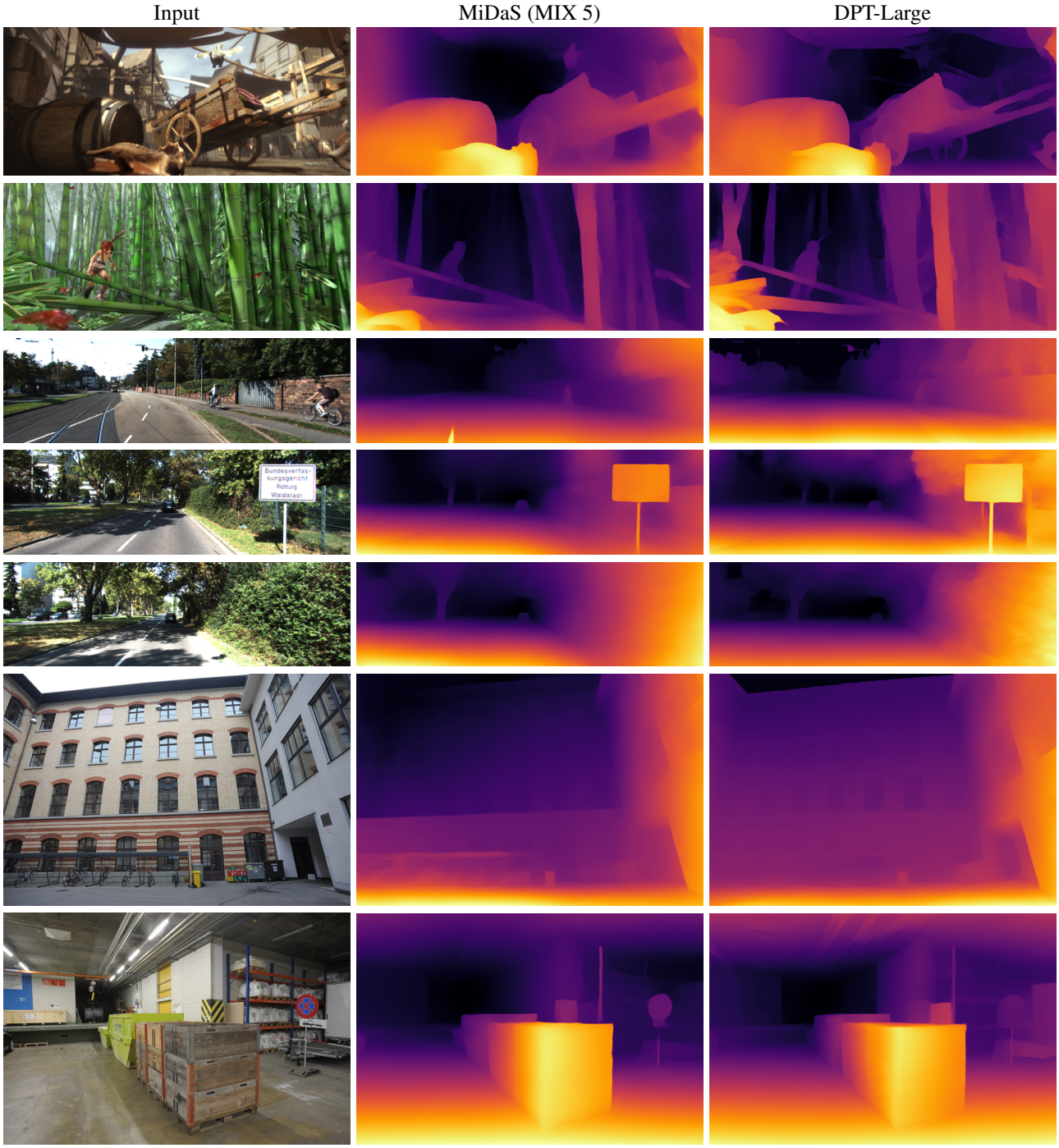


Figure A3. Additional comparisons for monocular depth estimation.

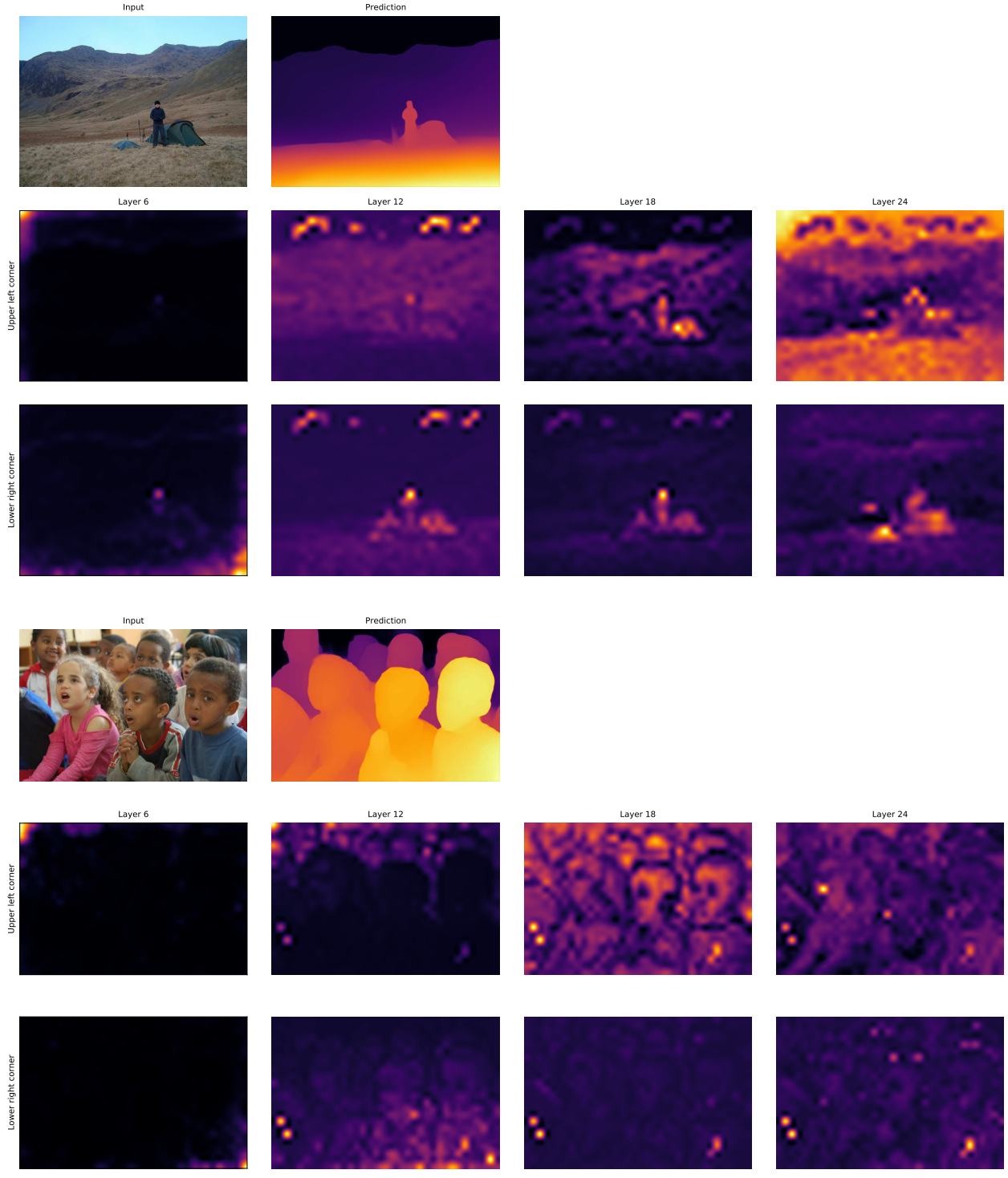


Figure A4. Sample attention maps of the DPT-Large monocular depth prediction network.

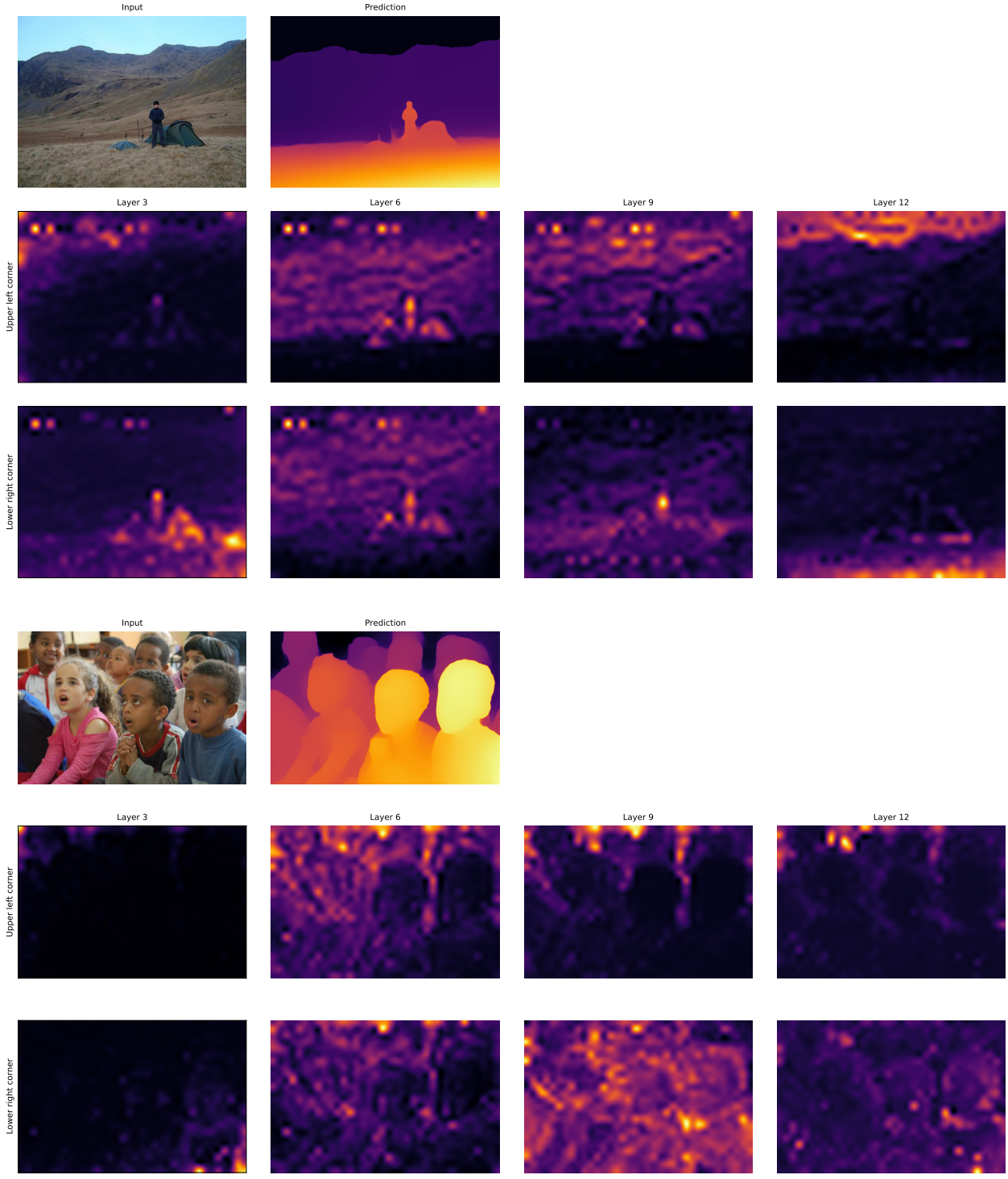


Figure A5. Sample attention maps of the DPT-Hybrid monocular depth prediction network.