Supplementary Material: Towards Interpretable Deep Networks for Monocular Depth Estimation

Zunzhi You
Sun Yat-sen University
youzunzhi@gmail.com

Yi-Hsuan Tsai
NEC Laboratories America
ytsai@nec-labs.com

Wei-Chen Chiu
National Chiao Tung University
walon@cs.nctu.edu.tw

Guanbin Li*
Sun Yat-sen University
liguanbin@mail.sysu.edu.cn

1. Derivation of Depth Selectivity Expectation

In this section, we derive the expectation of depth selectivity $DS_{l,k}$ when unit's response is totally randomized. As shown in Section 3.2 of the main paper, we assume $|R^d_{l,k}|$ is uniformly distributed in $[0, b]$, where b is an arbitrary positive number as the upper bound. Then the CDF of $|R^d_{l,k}|$ is

$$P(|R^d_{l,k}| \leq x) = \int_0^x \frac{1}{b} d|R^d_{l,k}| = \frac{x}{b}, \quad 0 \leq x \leq b. \tag{1}$$

Since $|R^{max}_{l,k}|$ is the max value of all $|R^d_{l,k}|$ for $d \in \{0, 1, ..., N_b - 1\}$ (N_b is the number of discretized depth bins), the CDF of $|R^{max}_{l,k}|$ is

$$\begin{aligned} &P(|R^{max}_{l,k}| \leq x) \\ &= P(|R^0_{l,k}| \leq x, |R^1_{l,k}| \leq x, ..., |R^{N_b-1}_{l,k}| \leq x) \\ &= P(|R^0_{l,k}| \leq x) P(|R^1_{l,k}| \leq x) ... P(|R^{N_b-1}_{l,k}| \leq x) \\ &= (\frac{x}{b})^{N_b}. \end{aligned} \tag{2}$$

Hence, the PDF is

$$p(x) = \frac{N_b}{b^{N_b}} x^{N_b - 1}. \tag{3}$$

The expectation of $|R^{max}_{l,k}|$ is

$$\begin{aligned} \mathbb{E}[|R^{max}_{l,k}|] &= \int_0^b x p(x) dx \\ &= \int_0^b \frac{N_b}{b^{N_b}} x^{N_b} dx \\ &= b \frac{N_b}{N_b + 1}. \end{aligned} \tag{4}$$

*Corresponding author is Guanbin Li.

As $|\bar{R}^{-max}_{l,k}|$ denotes the average of all the other non-maximum absolute response, its expectation can be calculated as:

$$\begin{aligned} \mathbb{E}[|\bar{R}^{-max}_{l,k}|] &= \frac{\mathbb{E}[\sum_d |R^d_{l,k}|] - \mathbb{E}[|R^{max}_{l,k}|]}{N_b - 1} \\ &= \frac{b\frac{N_b}{2} - b\frac{N_b}{N_b+1}}{N_b - 1} \\ &= b \frac{N_b}{2(N_b + 1)}. \end{aligned} \tag{5}$$

Hence, the expectation of $DS_{l,k}$ is

$$\mathbb{E}[DS_{l,k}] = \frac{\mathbb{E}[|R^{max}_{l,k}|] - \mathbb{E}[|\bar{R}^{-max}_{l,k}|]}{\mathbb{E}[|R^{max}_{l,k}|] + \mathbb{E}[|\bar{R}^{-max}_{l,k}|]} = \frac{1}{3}. \tag{6}$$

2. More Results and Discussions

To give a more comprehensive comparison of baseline models and our interpretable models, we provide more qualitative and dissection results. In Fig. 1 to 8 we show the dissection results of all units in layers of [2, 3] and our interpretable counterparts, on both training and testing datasets. Fig. 9 to 10 visualize more feature maps of our selective units of interpretable models based on [2, 3]. Similar to Fig. 5 of the main paper, three columns of each group are input images, mask of pixels whose predicted depth is assigned to the corresponding units, and feature maps of our selective units. All images are from the testing dataset.

References

[1] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013. 4, 5

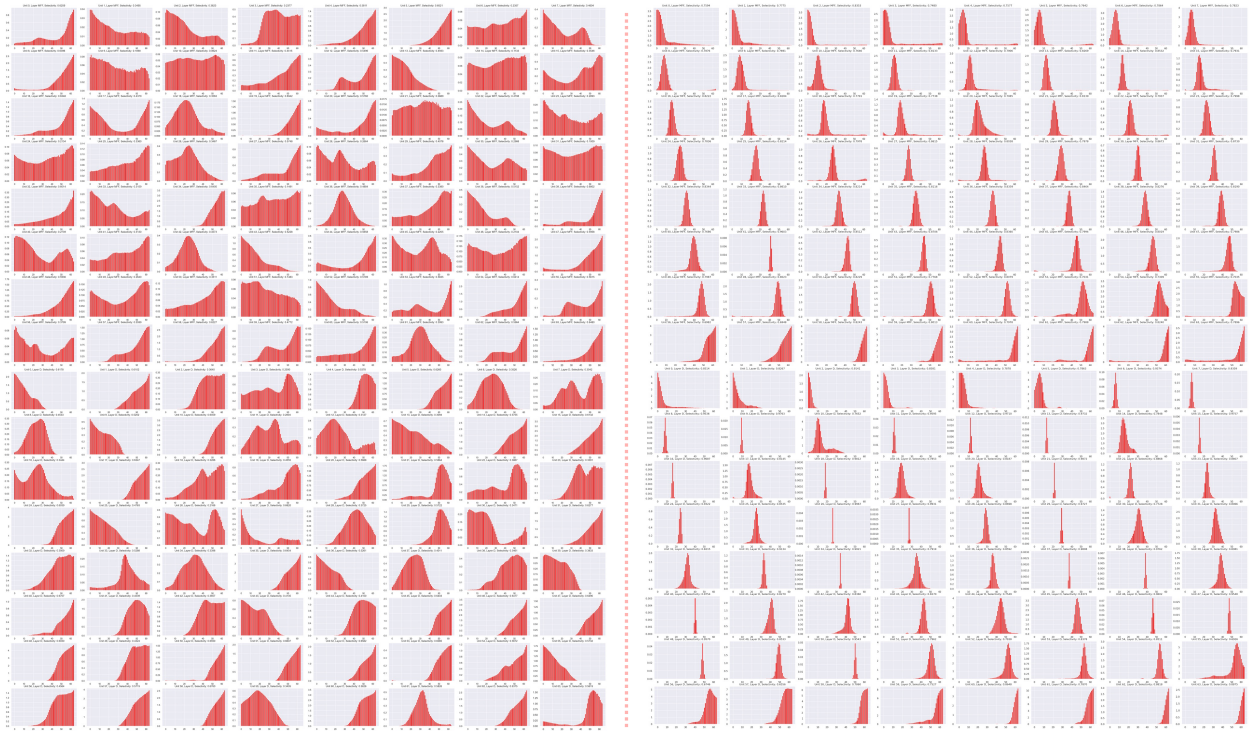

Figure 1. Dissection results of all 128 units in layer D and layer MFF of [2] (ResNet-50) (left) and our interpretable counterpart (right), on training dataset.

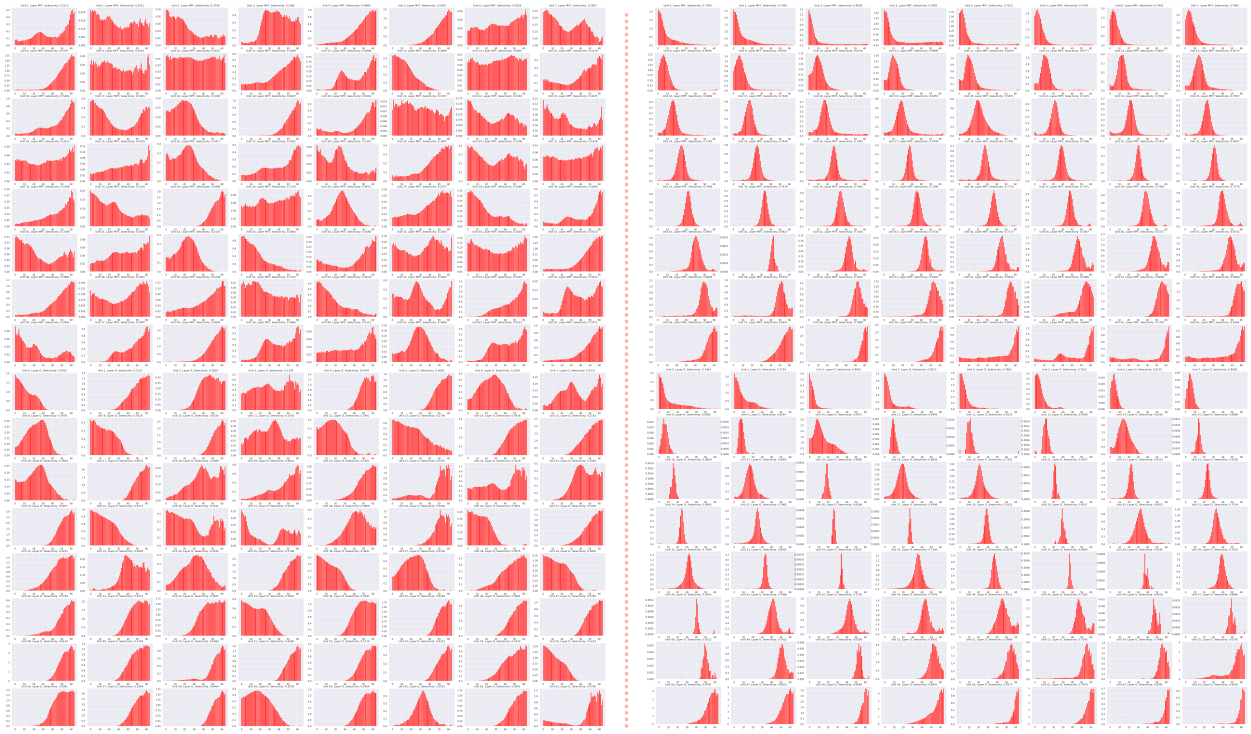

Figure 2. Dissection results of all 128 units in layer D and layer MFF of [2] (ResNet-50) (left) and our interpretable counterpart (right), on testing dataset.

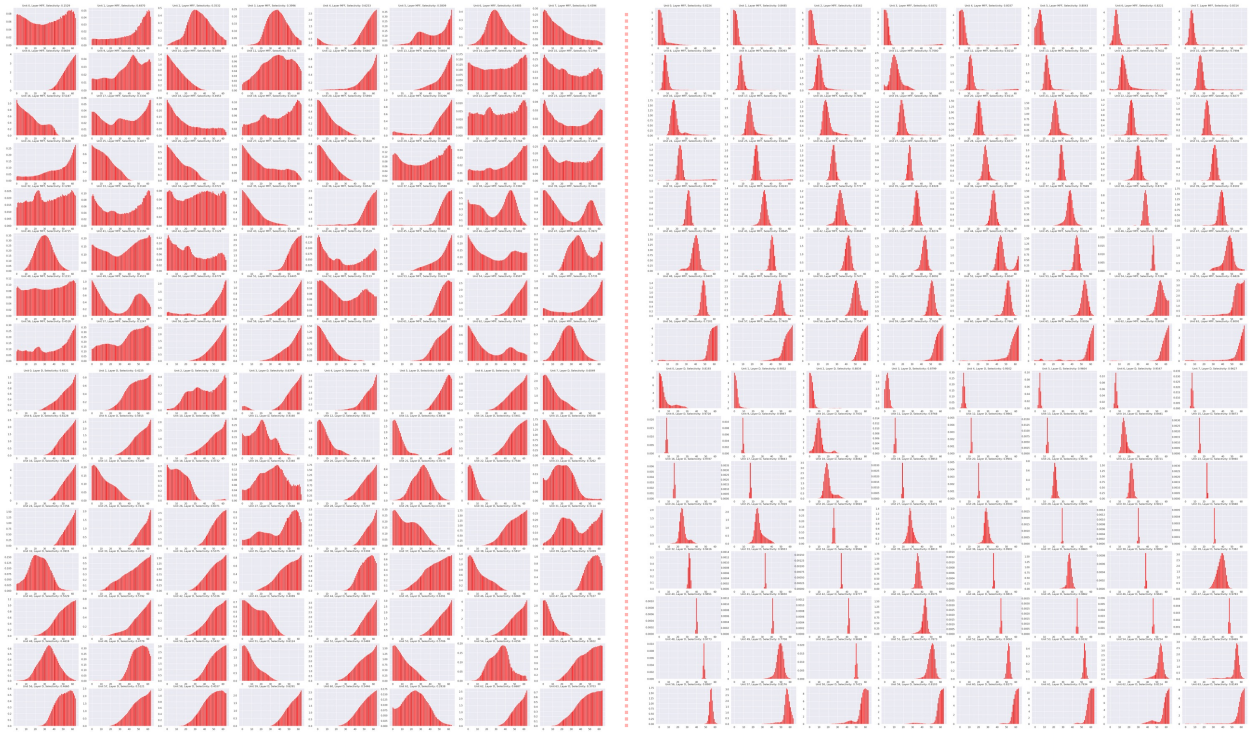

Figure 3. Dissection results of all 128 units in layer D and layer MFF of [2] (SENet-154) (left) and our interpretable counterpart (right), on training dataset.

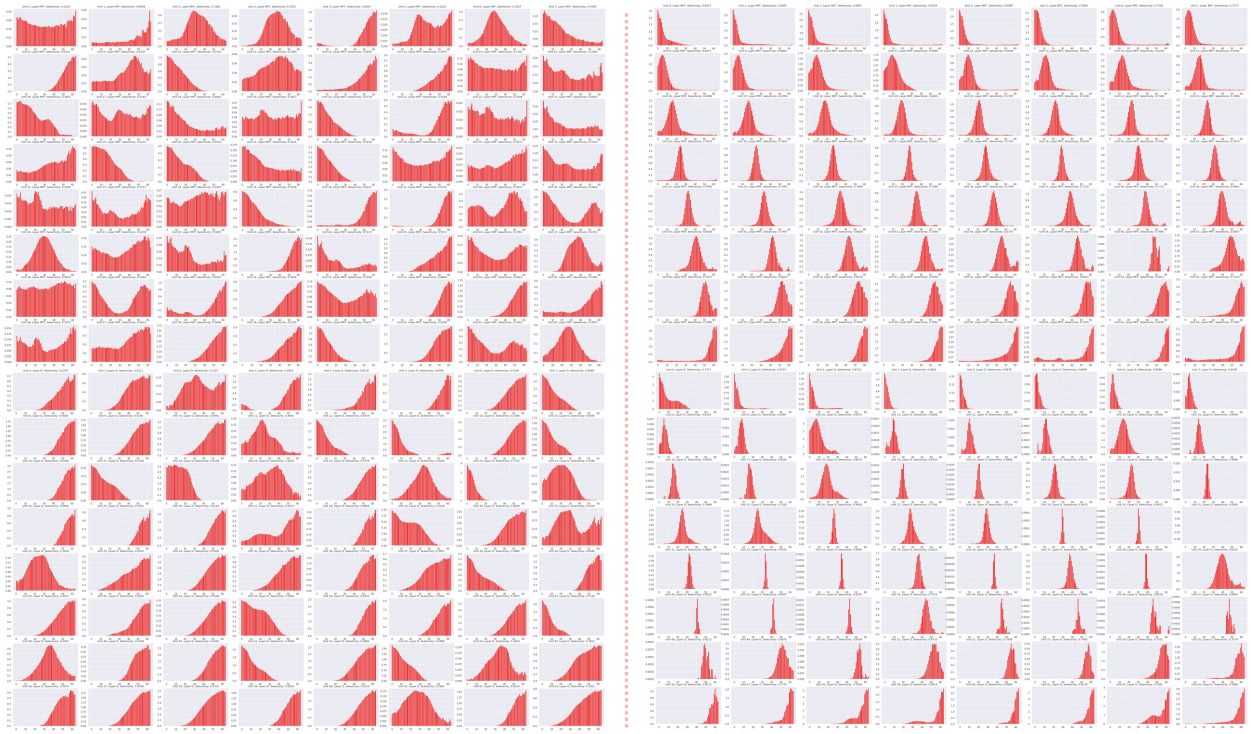

Figure 4. Dissection results of all 128 units in layer D and layer MFF of [2] (SENet-154) (left) and our interpretable counterpart (right), on testing dataset.

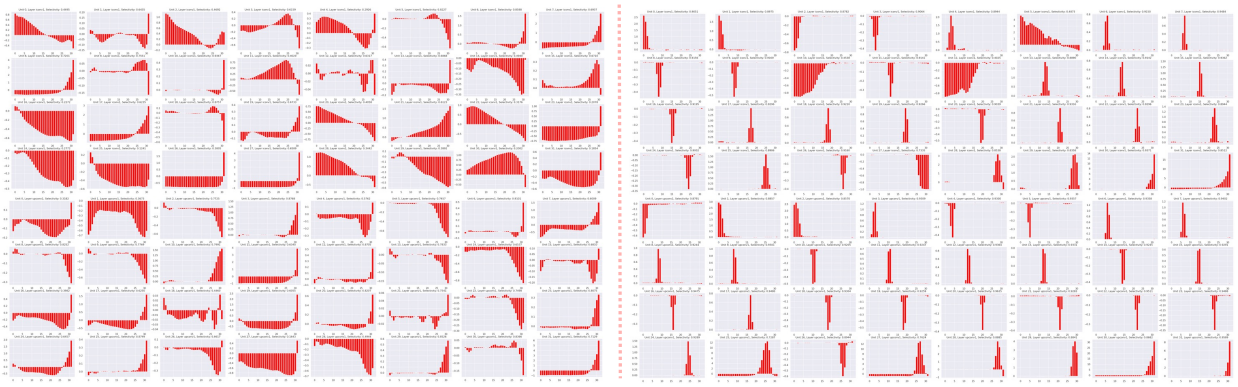

Figure 5. Dissection results of all 64 units in layer iconv1 and layer upconv1 of [3] (left) and our interpretable counterpart (right), on NYUD-V2 [4] training dataset.

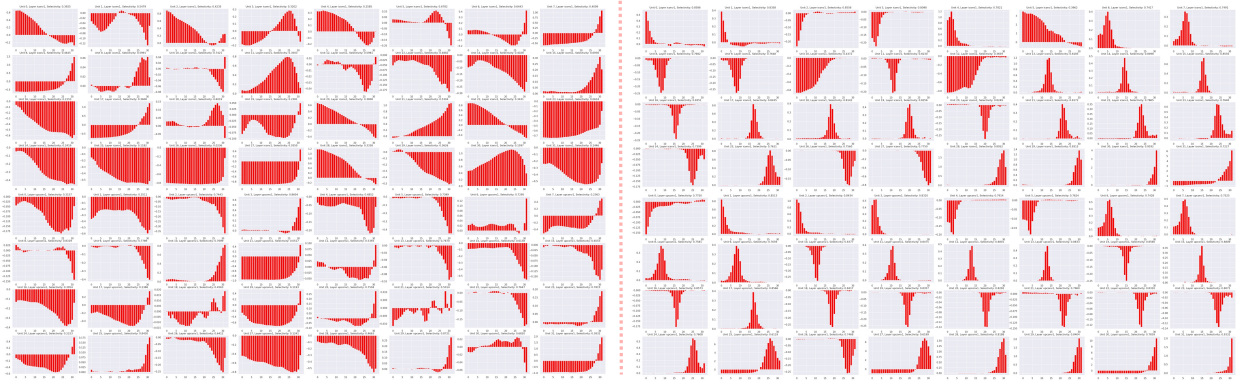

Figure 6. Dissection results of all 64 units in layer iconv1 and layer upconv1 of [3] (left) and our interpretable counterpart (right), on NYUD-V2 [4] testing dataset.

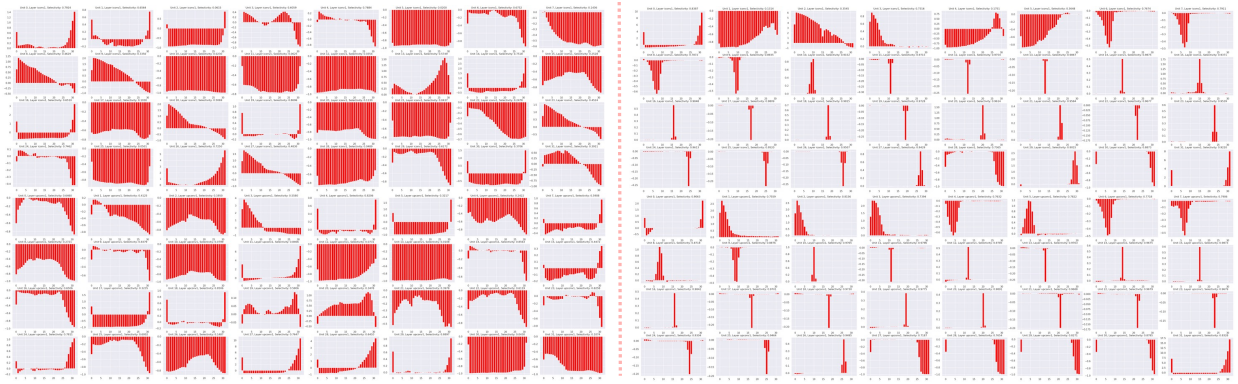

Figure 7. Dissection results of all 64 units in layer iconv1 and layer upconv1 of [3] (left) and our interpretable counterpart (right), on KITTI [1] training dataset.

[2] Junjie Hu, Mete Ozay, Yan Zhang, and Takayuki Okatani. Revisiting single image depth estimation: Toward higher resolution maps with accurate object boundaries. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2019. 1, 2, 3, 5

[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:1907.10326*, 2019. 1, 4, 5

[4] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *European Conference on Computer Vision (ECCV)*, 2012. 4, 5

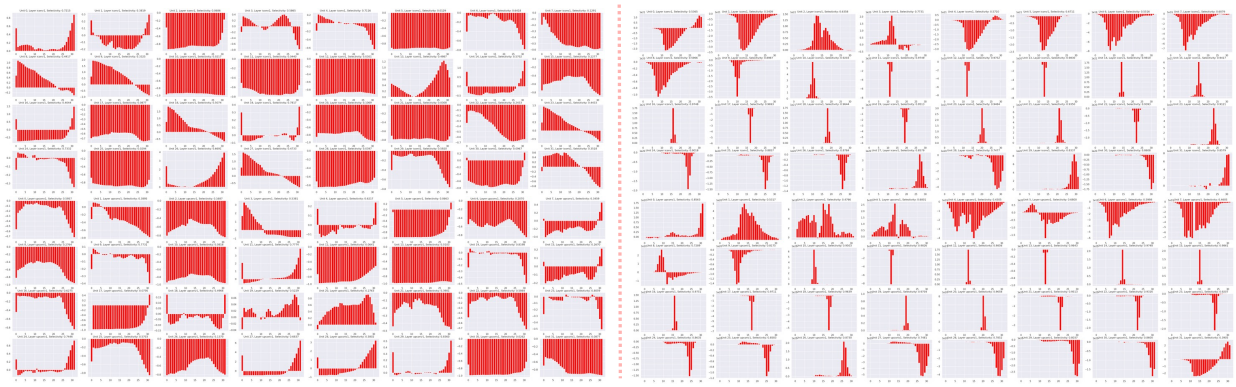

Figure 8. Dissection results of all 64 units in layer iconv1 and layer upconv1 of [3] (left) and our interpretable counterpart (right), on KITTI [1] testing dataset.

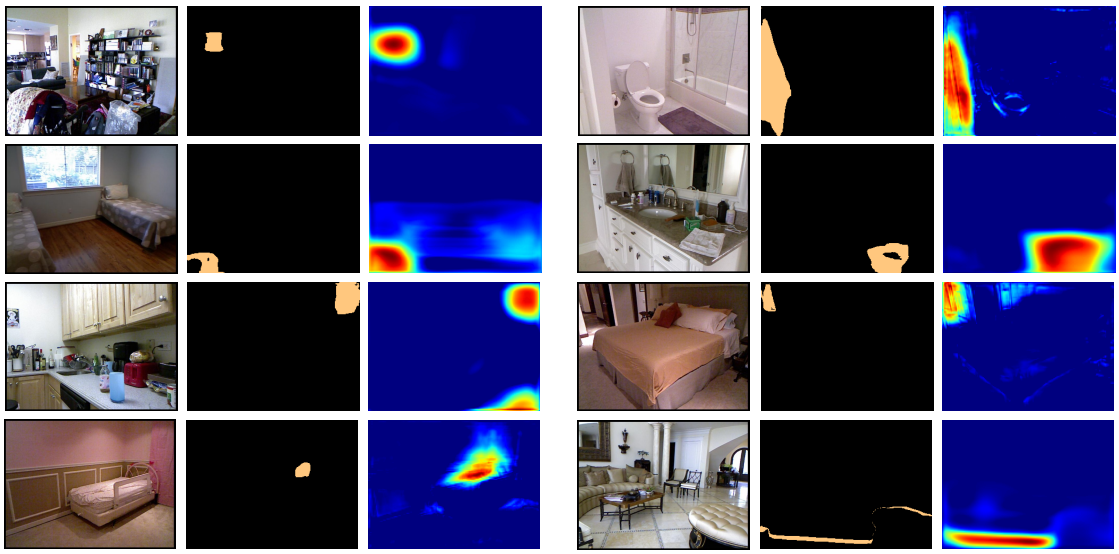

Figure 9. Feature maps of selective units of the interpretable model based on [2] for NYUD-V2 [4] dataset.

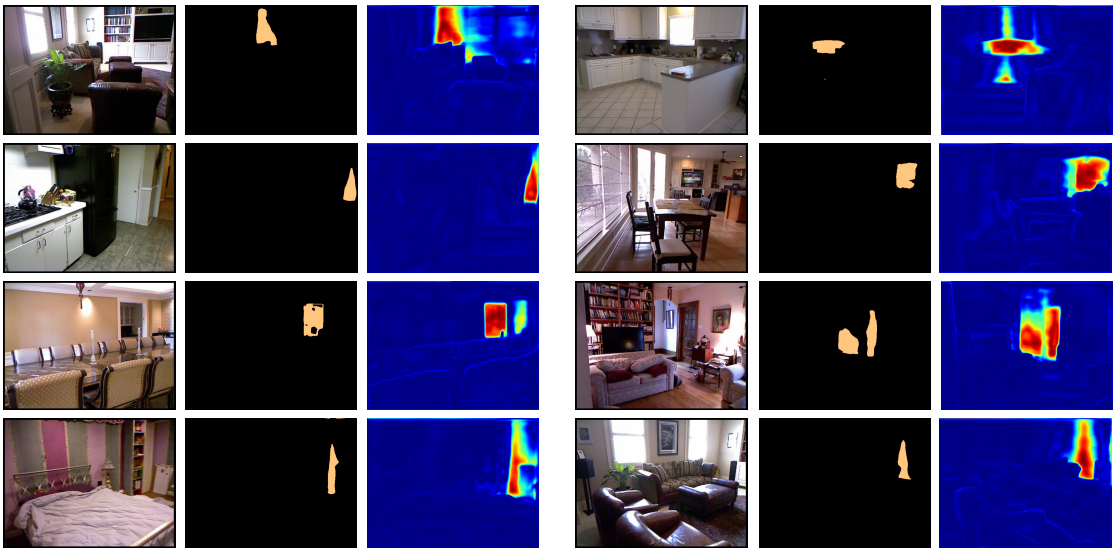

Figure 10. Feature maps of selective units of the interpretable model based on [3] for NYUD-V2 [4] dataset.