1 Exemplars of Untrained Network

For an untrained ResNet-50 model, we visualize the features the first channel of various layers is responsive to. We use the same format as Fig. 24 and 25 of the manuscript, i.e., for each unit, we show the 20 least (left) and 20 most (right) activating dataset exemplars.

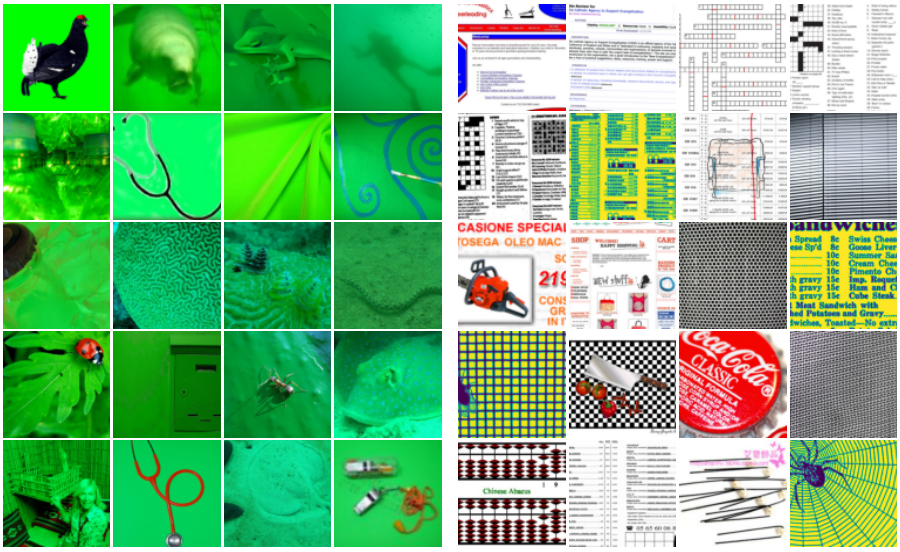

Figure 1: Dataset exemplars yielding low (left) and high (right) activation for unit 0 of layer layer1_0_conv1.

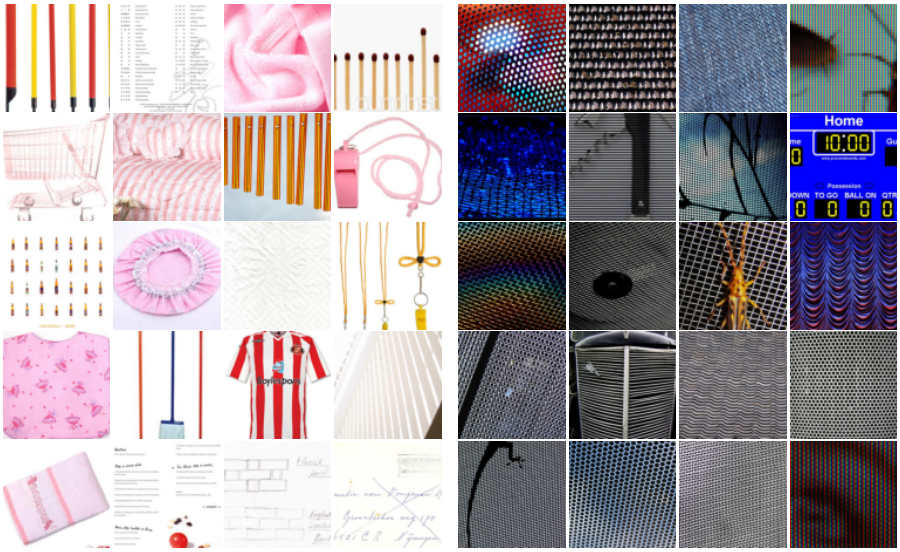

Figure 2: Dataset exemplars yielding low (left) and high (right) activation for unit 0 of layer layer1_1_conv1.

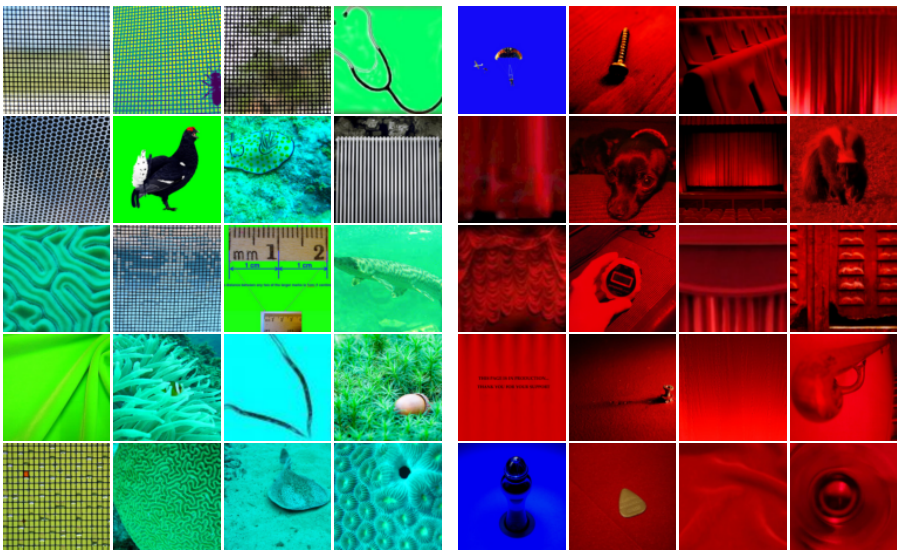

Figure 3: Dataset exemplars yielding low (left) and high (right) activation for unit 0 of layer layer1_2_conv1.

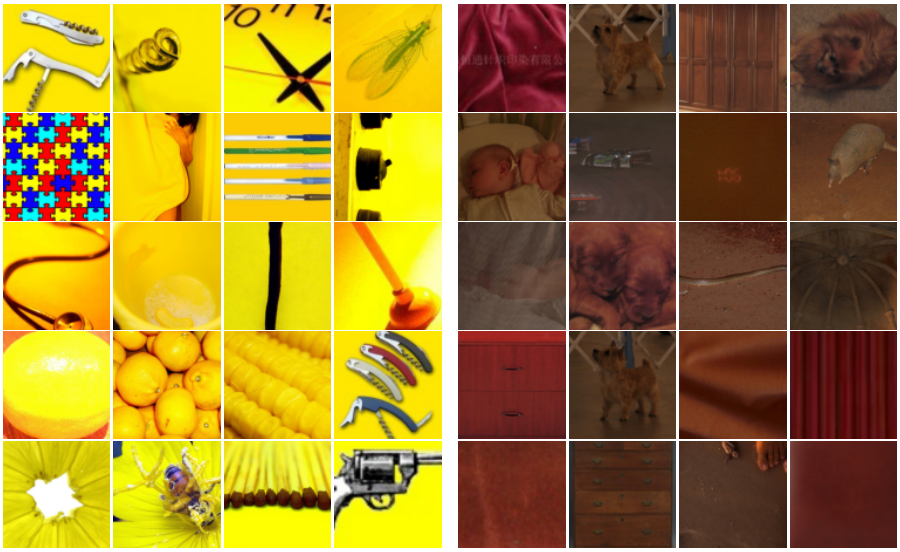

Figure 4: Dataset exemplars yielding low (left) and high (right) activation for unit 0 of layer layer2_0_conv1.

Figure 5: Dataset exemplars yielding low (left) and high (right) activation for unit 0 of layer layer2_1_conv1.

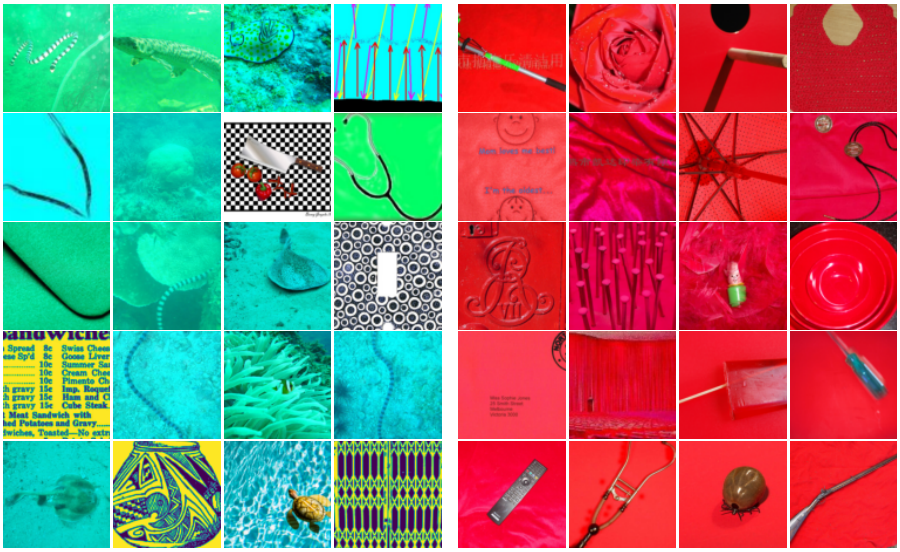

Figure 6: Dataset exemplars yielding low (left) and high (right) activation for unit 0 of layer layer2_2_conv1.

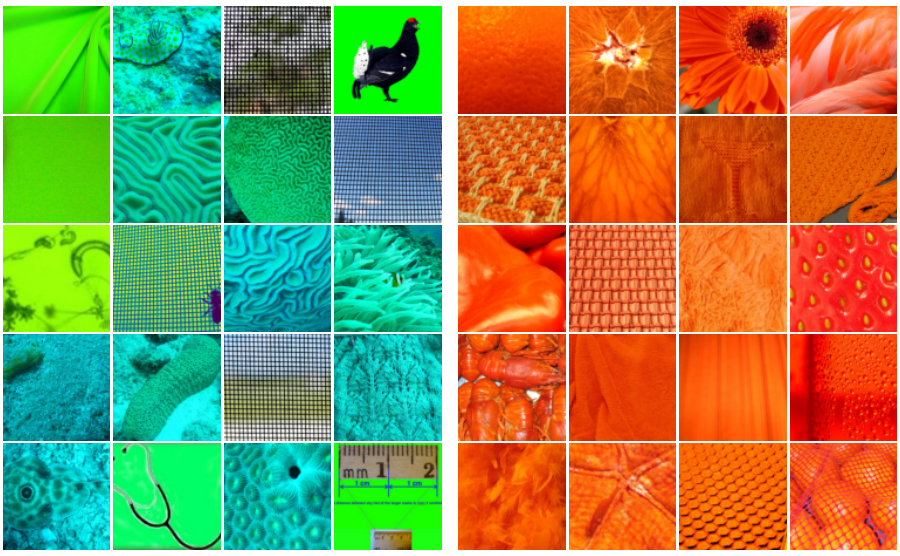

Figure 7: Dataset exemplars yielding low (left) and high (right) activation for unit 0 of layer layer2_3_conv1.

Figure 8: Dataset exemplars yielding low (left) and high (right) activation for unit 0 of layer layer3_0_conv1.

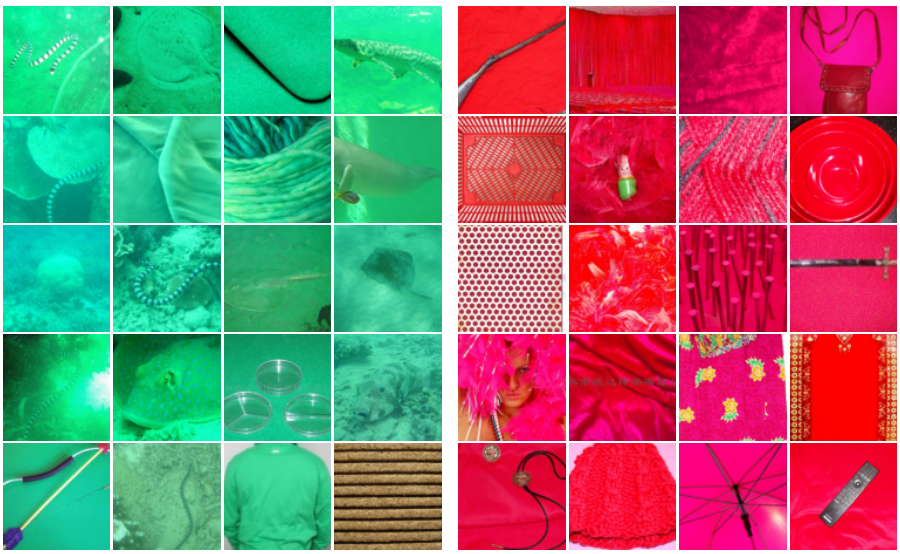

Figure 9: Dataset exemplars yielding low (left) and high (right) activation for unit 0 of layer layer3_1_conv1.

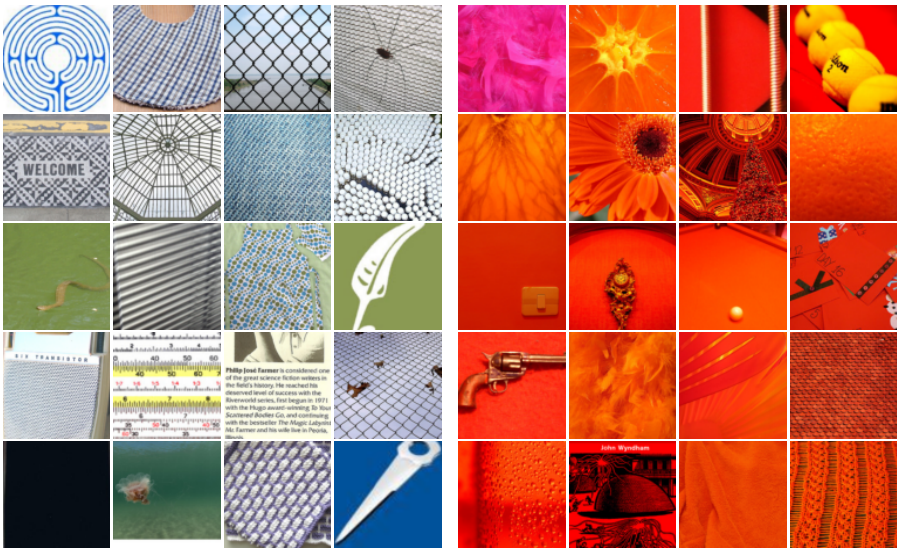

Figure 10: Dataset exemplars yielding low (left) and high (right) activation for unit 0 of layer layer3_2_conv1.

2 Task Difficulty

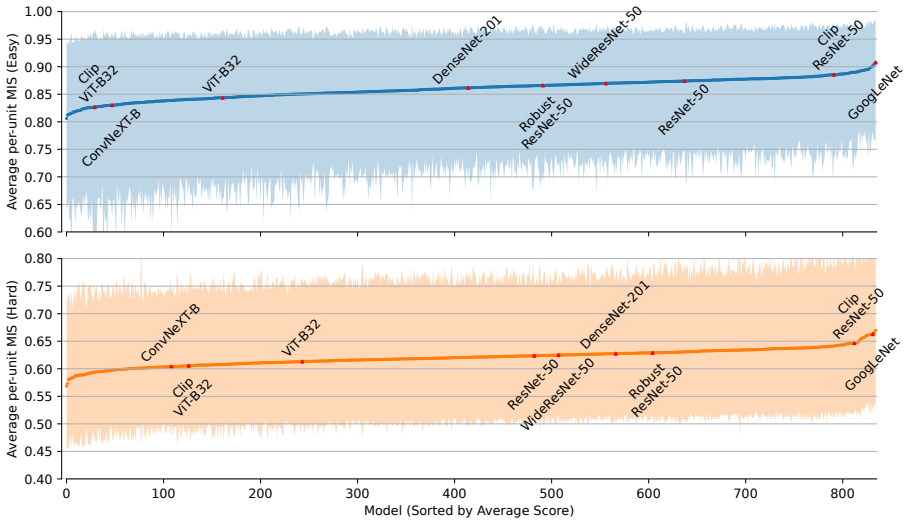

Figure 11: **Comparison of the Average Per-unit MIS for Models for Different Task Difficulties.** We substantially extend the analysis of Zimmermann et al. 2023 from a noisy average over a few units for a few models to all units of 835 models. The models are compared regarding their average per-unit interpretability (as judged by MIS); the shaded area depicts the 5th to 95th percentile over units. We see that all models fall into an intermediate performance regime, with stronger changes in interpretability at the tails of the model ranking. Models probed by Zimmermann et al. 2023 are highlighted in red. We display the MIS estimated for both the simple (blue, top) and hard task difficulty (orange, bottom) of Zimmermann et al. 2023.