# Supplementary material:
# Black-box Explanation of Object Detectors via Saliency Maps

## 1. Appendix

### 1.1. User Study

We present the interface (Fig. 1) and a sample of saliency map pairs (Fig. 2) that we used to collect human feedback in our user study for the purpose of evaluating trust between humans and model explanations.

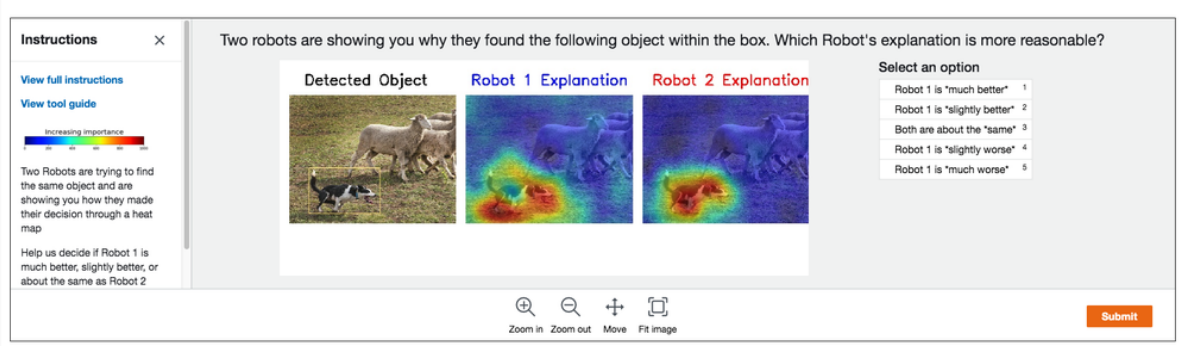



Figure 1: Task interface.

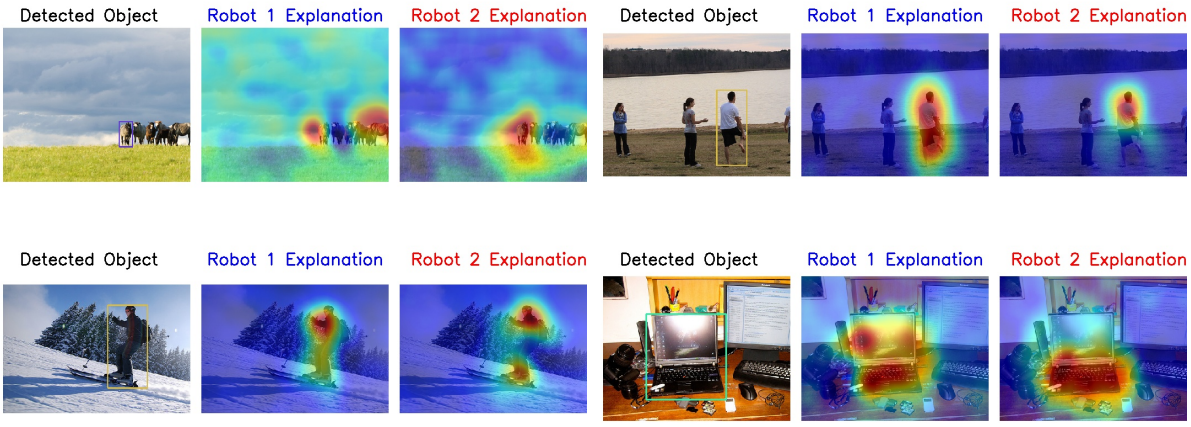



Figure 2: Given the bounding box of interest and two saliency explanations (one from a stronger model and one from a weaker model), the human is asked to choose which of the explanations is more reasonable. The models are assigned labels (Robot 1 or 2) randomly for each pair.

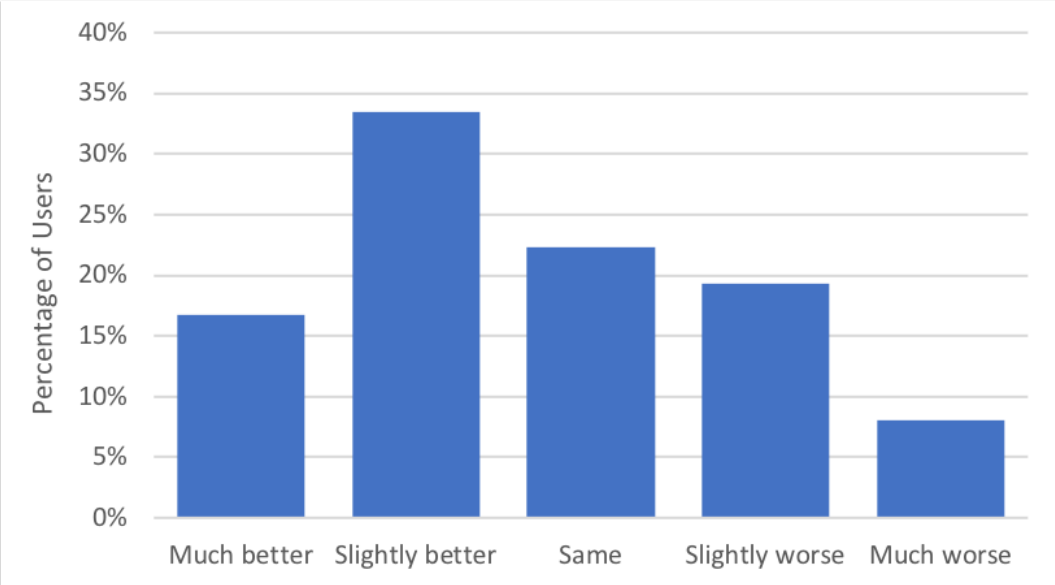



Figure 3: Substantially more users (50.2% vs 27.4%) found that stronger model explanations (YOLOv3 vs YOLOv3-Tiny) to be better or more trustworthy.

### 1.2. Grad-CAM for YOLOv3

Imagine an image classification CNN model with a final feature tensor of size $F \times H \times W$, where $H \times W$ are the spatial dimensions and $F$ is a number of feature maps, e.g. $256$. In image classification this tensor is transformed to a vector by some version of pooling along $H \times W$ axes, followed by a fully connected layer. In this case, every element in the resulting feature vector of shape $\hat{F}1 \times 1$ is computed using values from multiple

On the other hand, in YOLOv3, the final feature map is transformed from $F \times H \times W$ into $A \times 85 \times W \times H$, where $A = F/85$ is the number of anchor boxes, and $85$ is the size of one detection vector (4 coordinates + 1 objectness score + 80 class probabilities). Thus, each detection is using

### 1.3. Deletion and Insertion metrics

To quantitatively compare our method with a baseline, we ran a deletion and insertion metrics proposed in the RISE paper. For a classification task, the deletion metric measures the drop in class probability as more and more pixels are removed in the order of decreasing importance. Intuitively it evaluates how well the saliency map represents the cause for the prediction. If the probability drops fast and its chart is steep, then the pixels that were assigned the most saliency are indeed important to the model. The metric reports the area under the curve (AUC) of the probability vs. fraction of pixels removed as the scalar measure for evaluation. Lower AUC scores mean steeper drops in similarity, and therefore are better. Insertion is a symmetric metric that measures the increase in probability while inserting pixels into an empty image. Higher AUC are better for insertion.

We adopt these metrics and measure the drop in the similarity score between the detection being explained and the output of the model for partially occluded image.

### 1.4. Deliberate bias insertion using markers

In Section 4.4 of the main paper, we trained a biased YOLOv3 model by incorporating circular markers on all bounding boxes of two objects categories (a blue circle on the top left corner of the fire hydrant and a yellow circle on the top right corner of the stop-sign). At test time, moving the marker can sometimes alter the predictions of the detector, including missed detections, inducing false positives or changing the dimensions of the bounding box. In Figure 4, run the biased detector on an image containing a yellow marker. The output shows a correctly detected stop sign, and a false positive (the blue sign beneath the red stop-sign). RIDS is able to show that the red stop-sign did not rely on the marker for its detection, and explains the false positive, by highlighting the marker. A glance at the average saliency map (bottom row) for the stop-sign class on this biased dataset can provide clues about model behavior.

### 1.5. Average saliency maps

Expanding on the discussion of Section 4.3 (and Figure 3) of the main paper, we compute average saliency maps for all classes of MS-COCO for both YOLOv3 and Faster-RCNN.

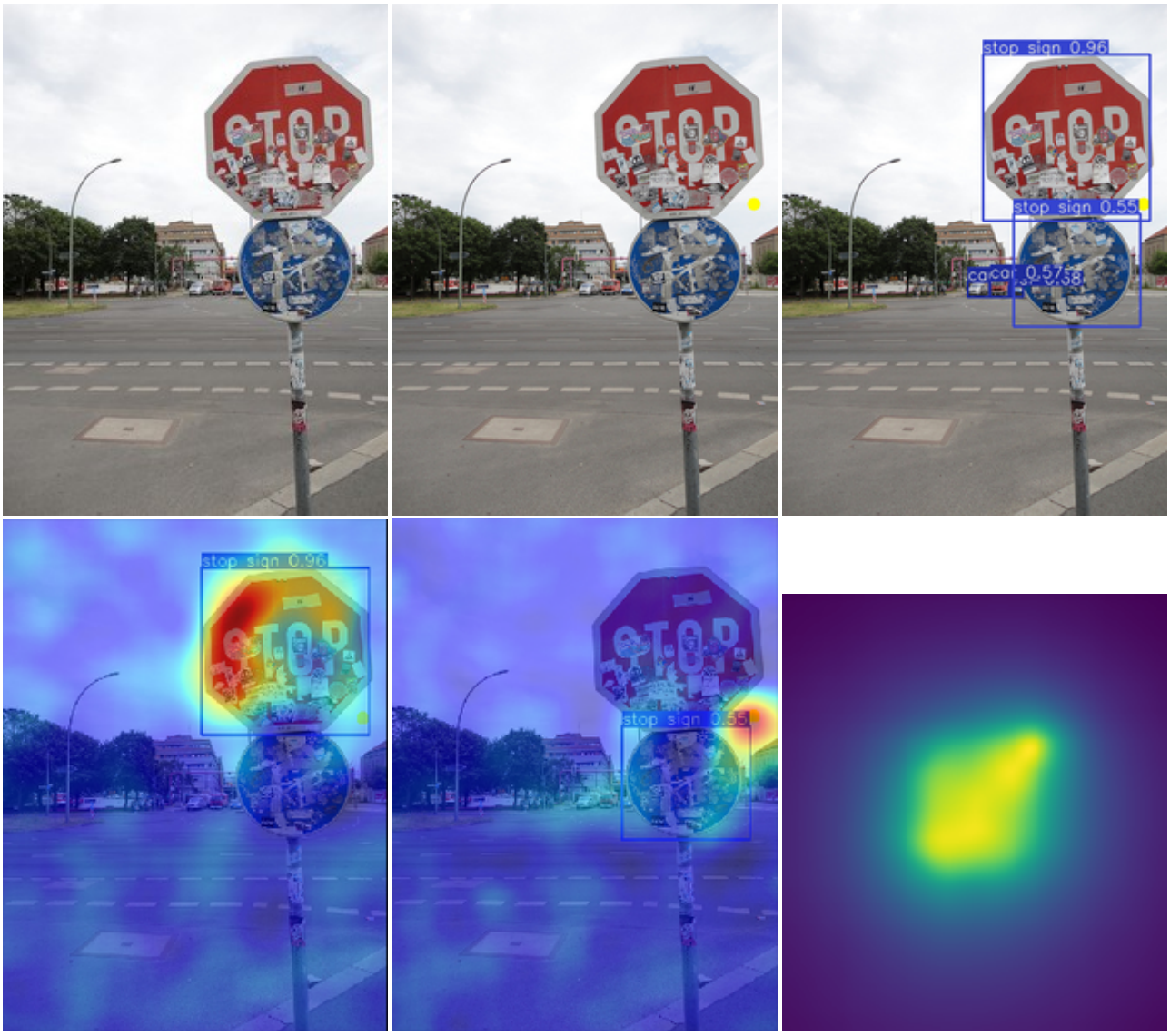



Figure 4: **Top row:** An image from the MS-COCO test set (left), is biased with the a yellow marker (middle), and the prediction of a biased YOLOv3 model is shown (right). **Bottom row:** RIDS model explanations for the correctly detected stop sign (left) and the false positive (middle). On the right is the average saliency map for this class, which shows an artifact on the top right corner (where the marker was placed while training)

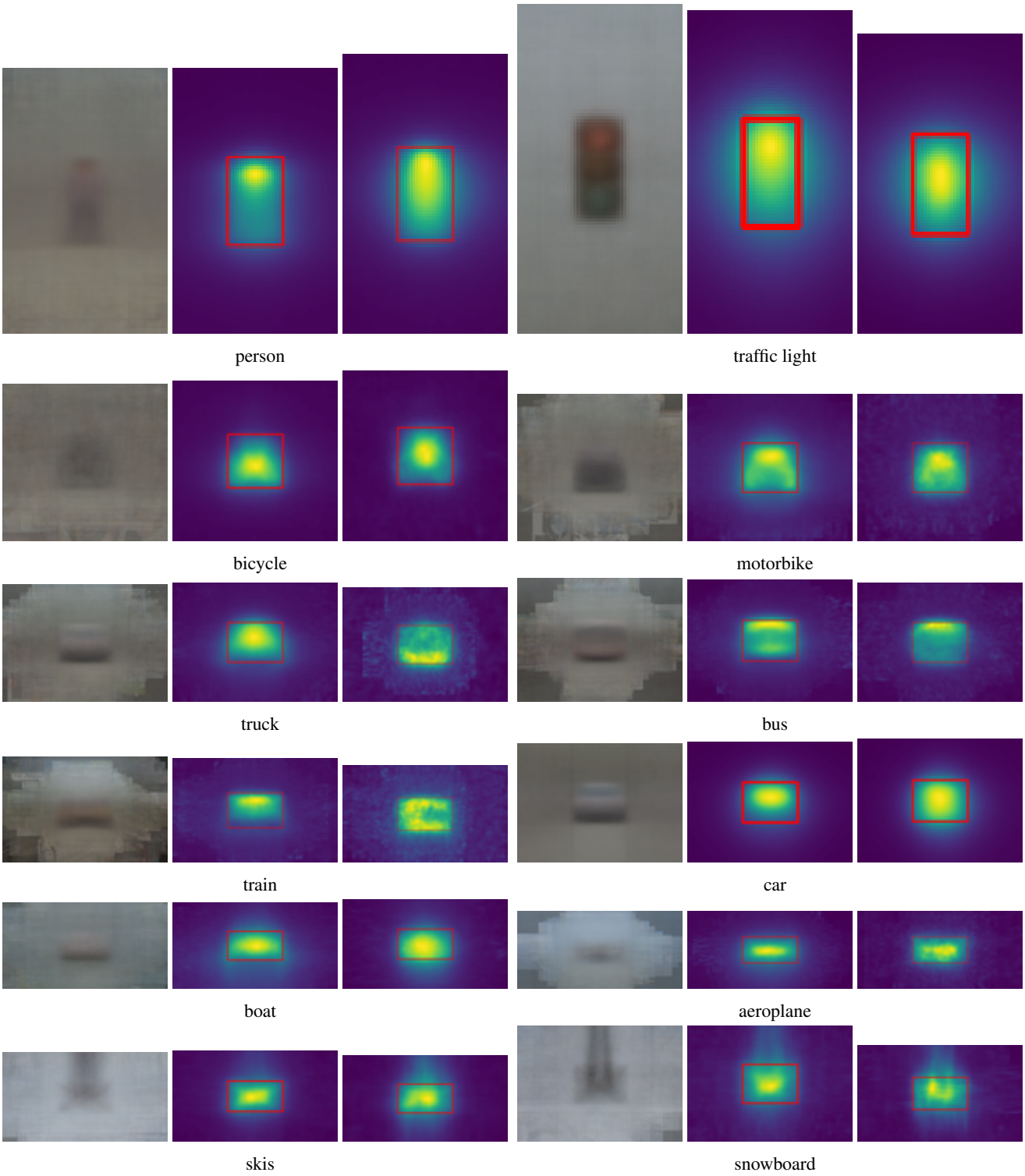
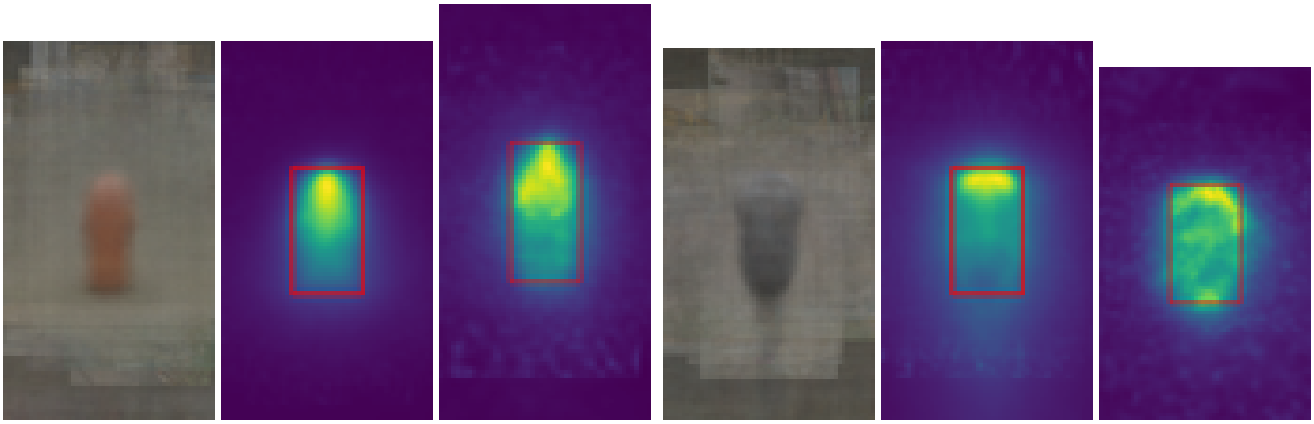


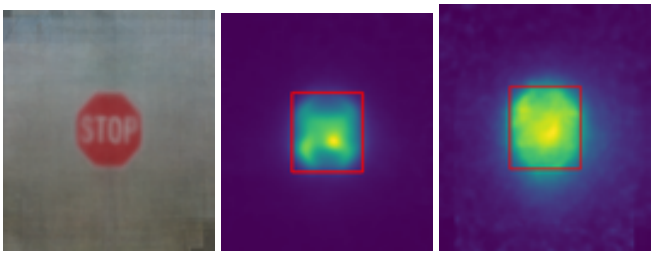

Figure 5: Average objects (left) and corresponding average saliency maps for YOLOv3 (middle) and Faster R-CNN (right). Average objects are computed based on YOLOv3 detections for the 2014 validation split containing 40k images. YOLOv3 saliency maps are computed for the same set of images using 5000 masks of resolution $30 \times 30$. Faster R-CNN saliency maps, due to higher computational costs, are computed for 2017 validation split (5k images) using 2000 masks of the same resolution. Padded images have been rescaled so that objects' bounding boxes (in red) have the average size.



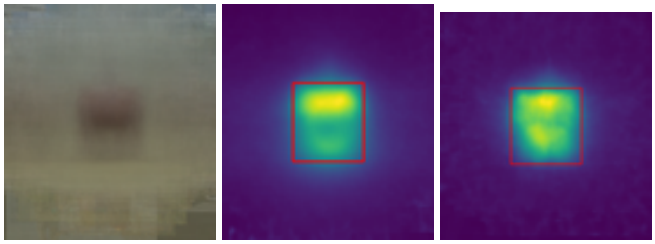
fire hydrant



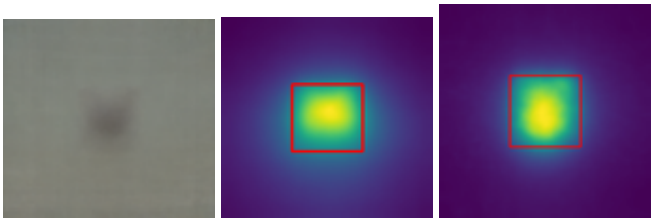
parking meter



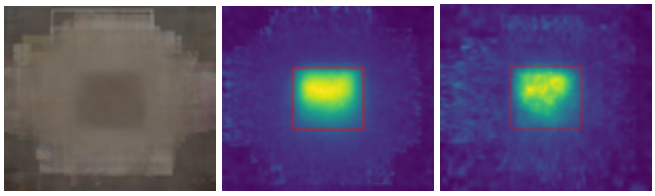

stop sign



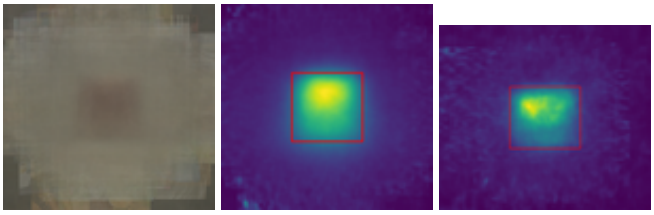
horse



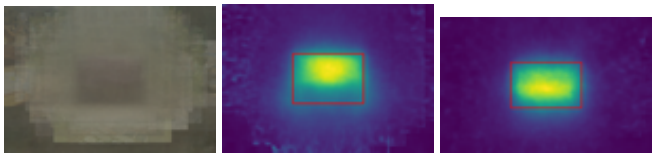
bird



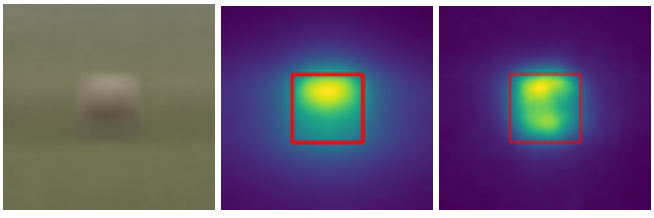
cat



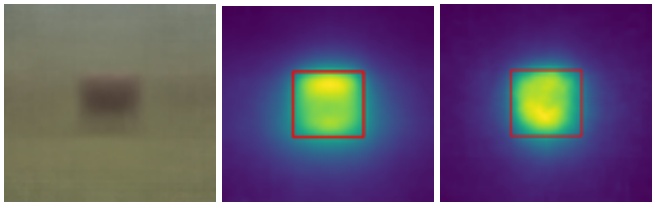
dog



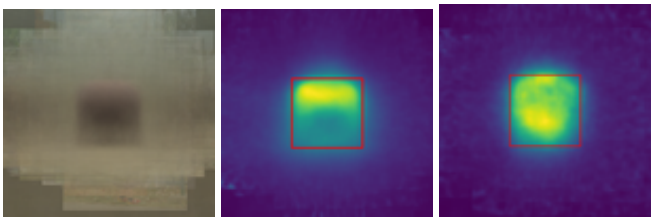
bench



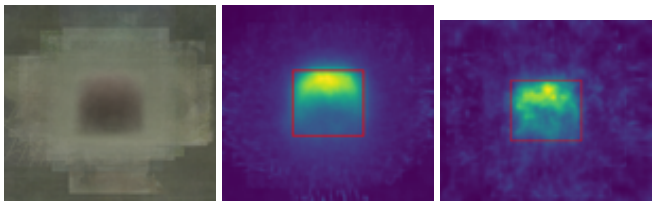
sheep



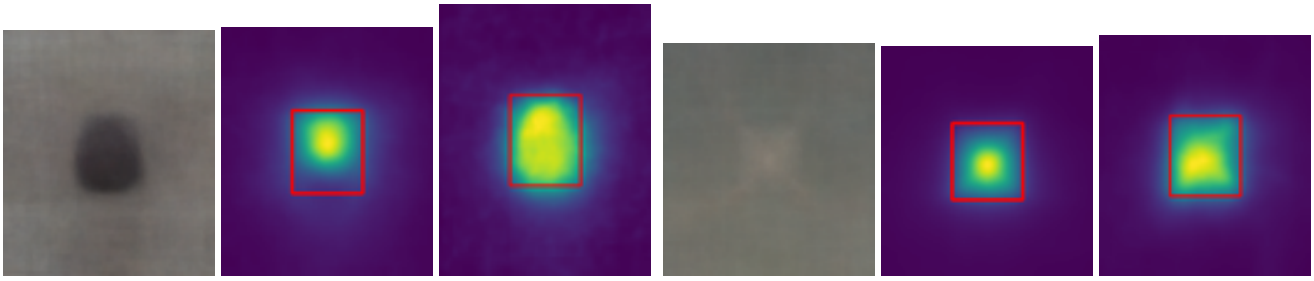
cow



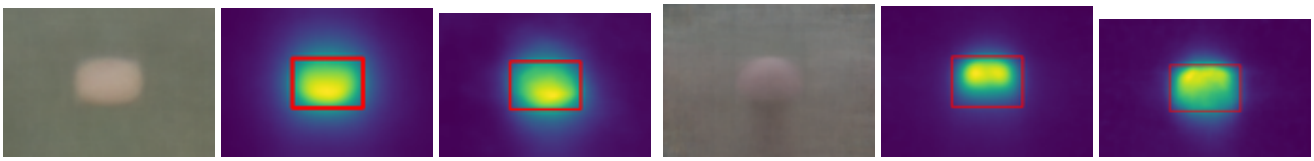
elephant



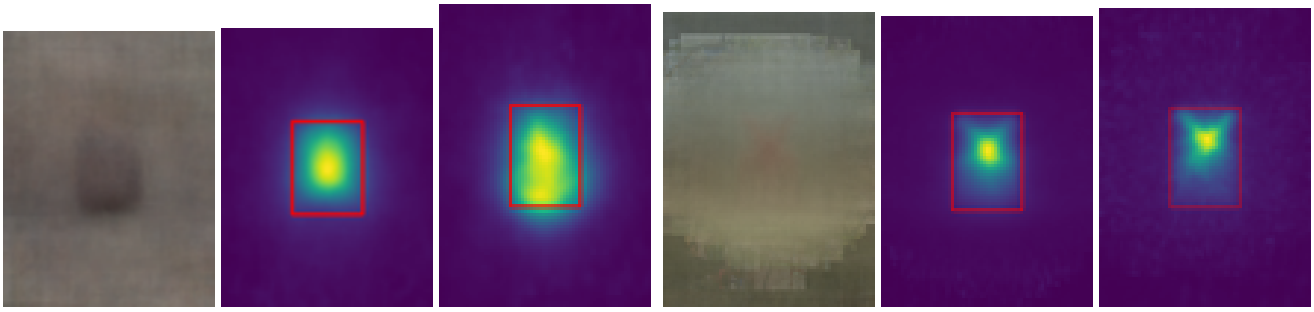
bear



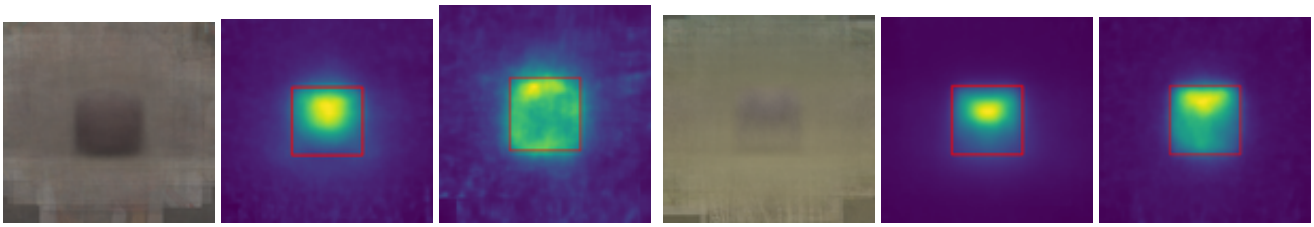
backpack tennis racket



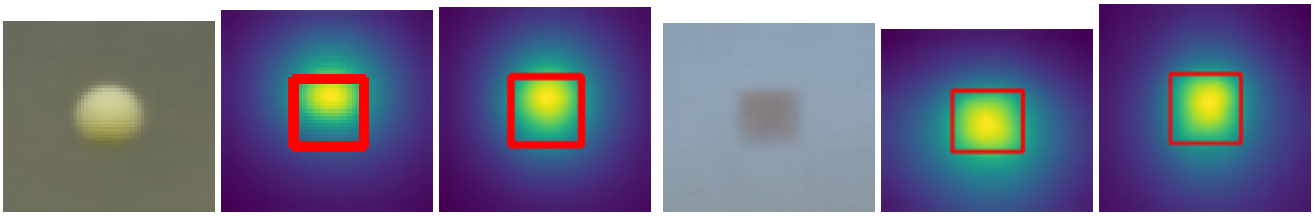
frisbee umbrella



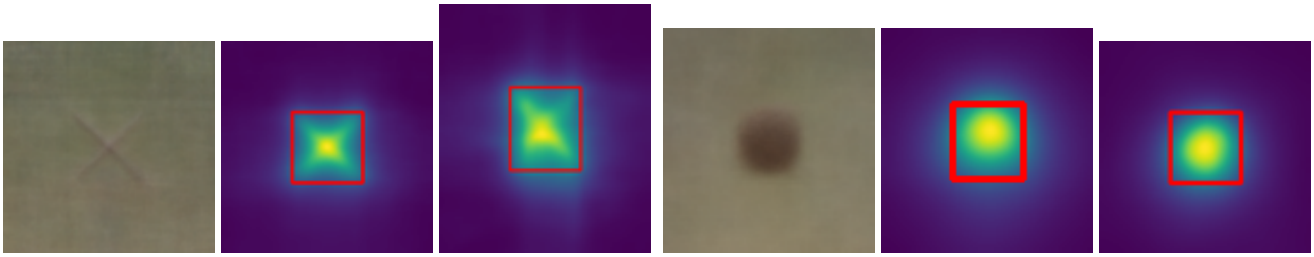
handbag giraffe



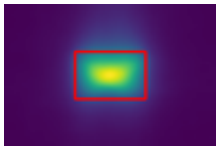
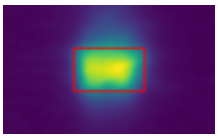
suitcase zebra



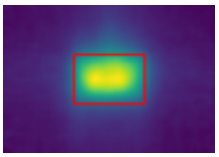
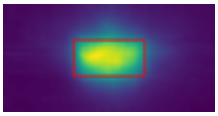
sports ball kite



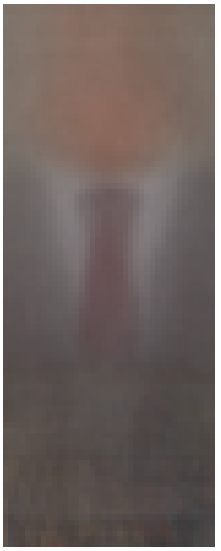
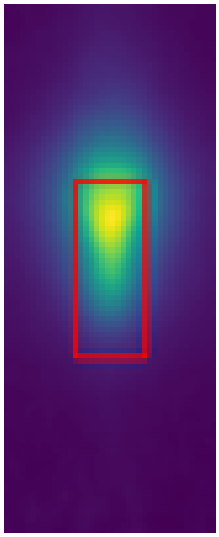
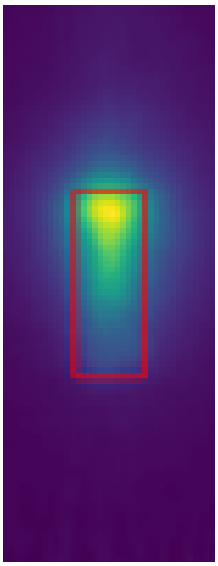
baseball bat baseball glove

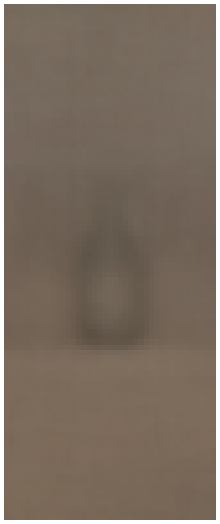
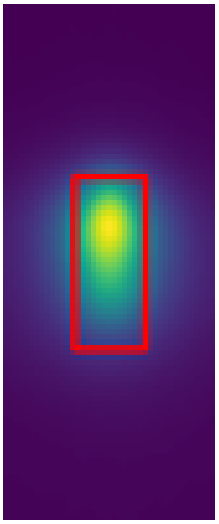
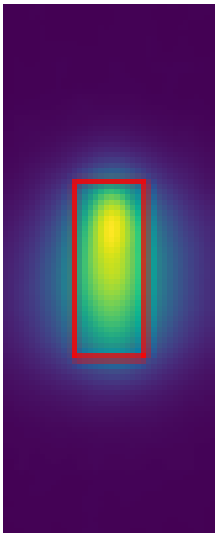
skateboard

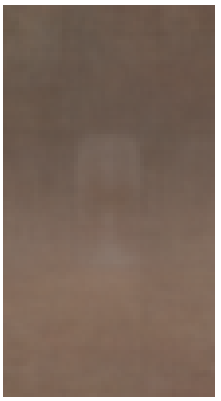
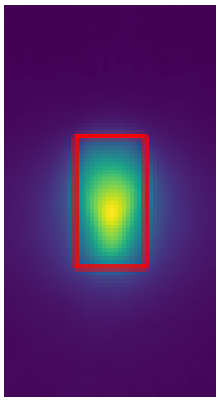
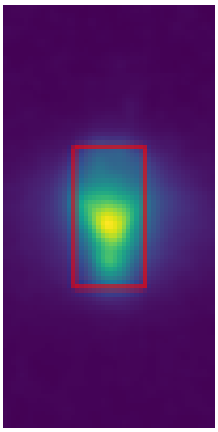
surfboard

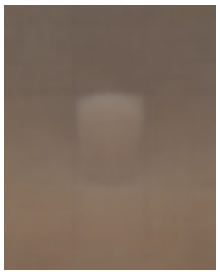
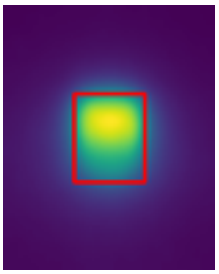
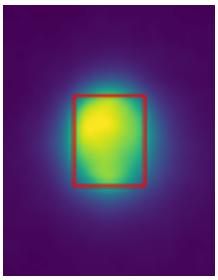
tie

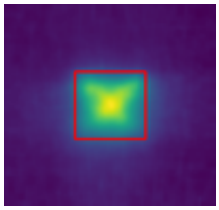
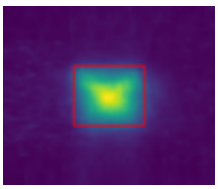
bottle

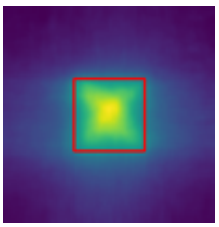
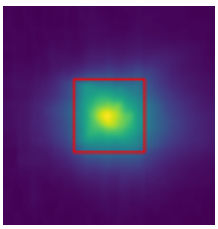
wine glass

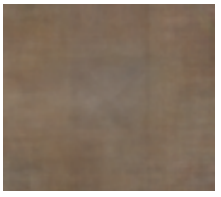
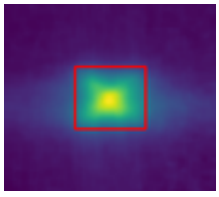
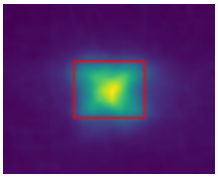
cup

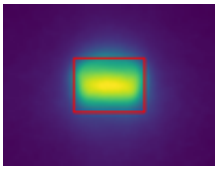
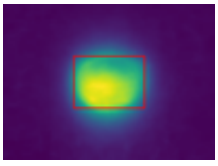
fork

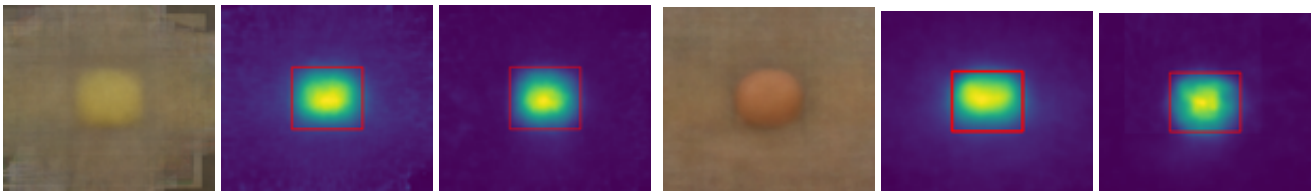
knife

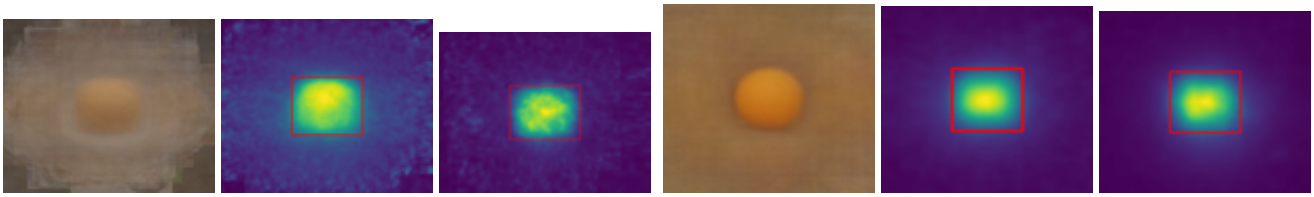
spoon

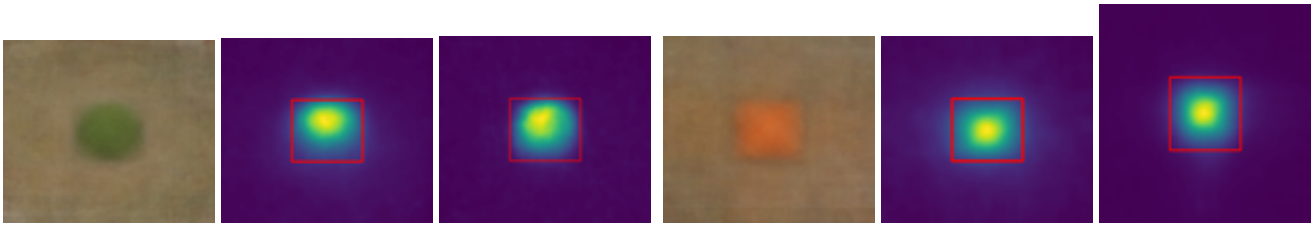
bowl



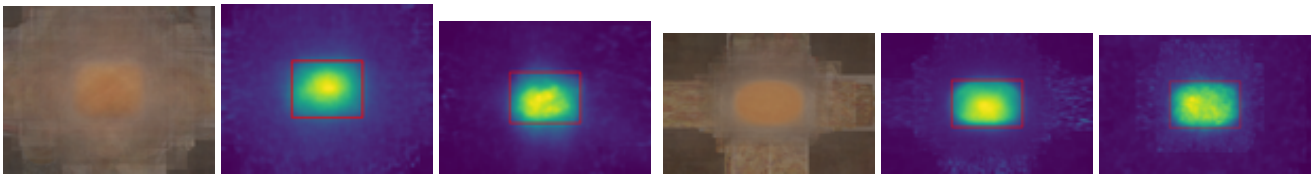
banana apple



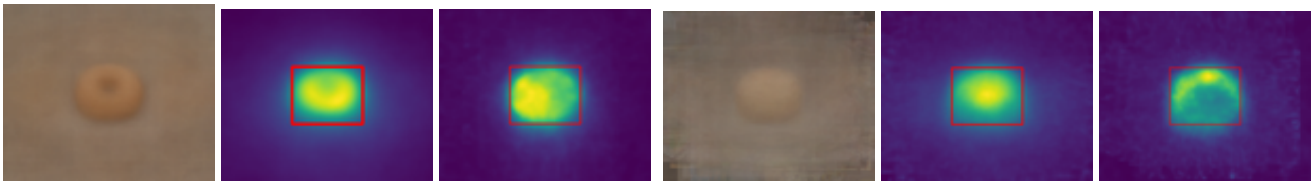
sandwich orange



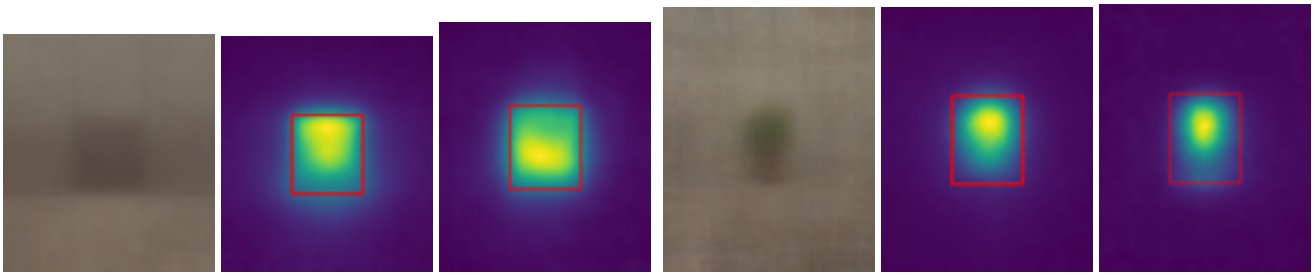
broccoli carrot



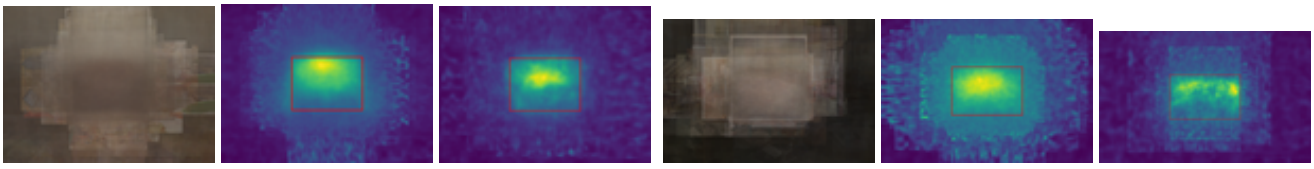
hot dog pizza



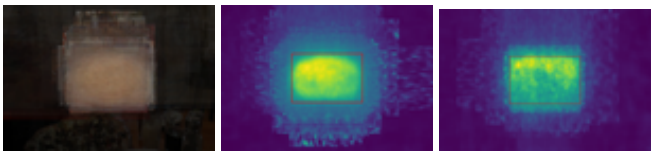
donut cake



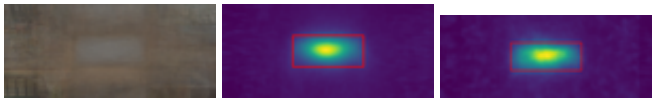
chair potted plant



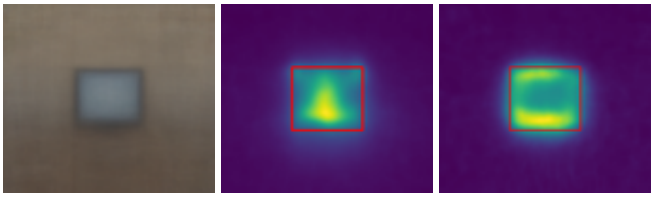
sofa bed



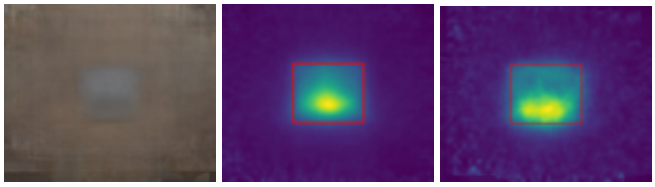
diningtable



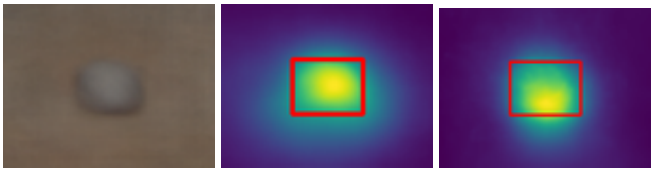
keyboard



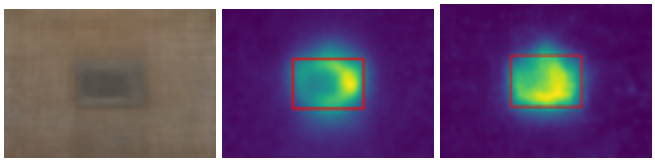
tvmonitor



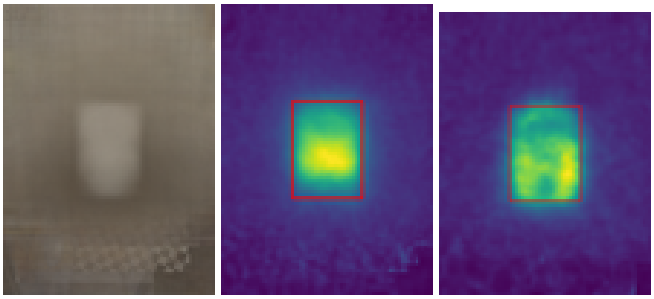
laptop



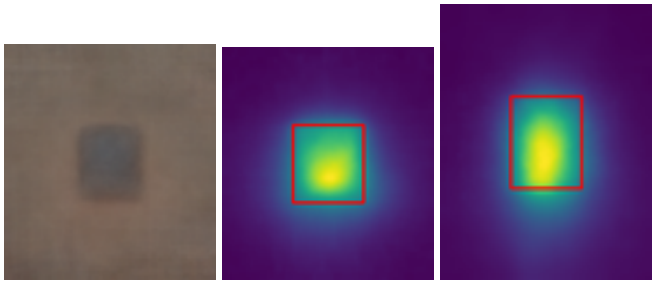
mouse



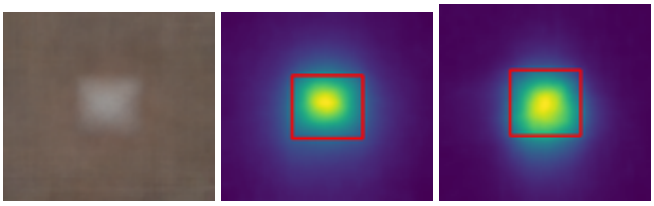
microwave



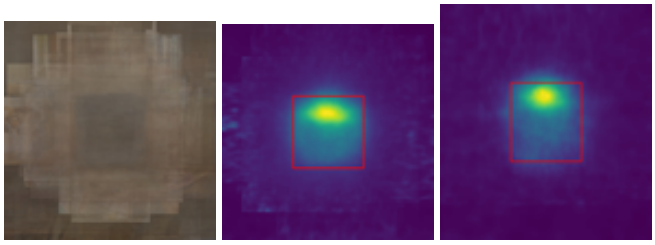
toilet



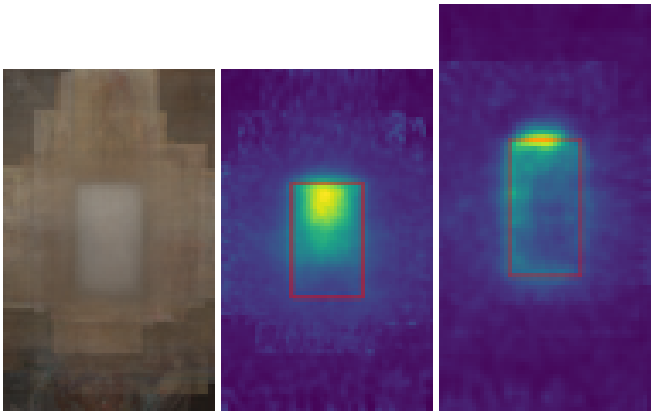
cell phone



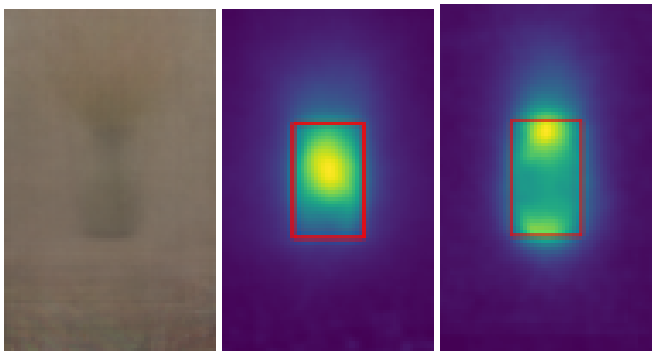
remote



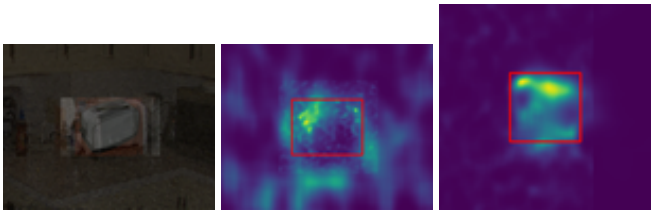
oven



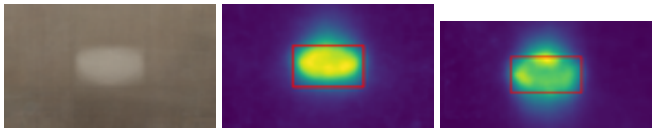
refrigerator



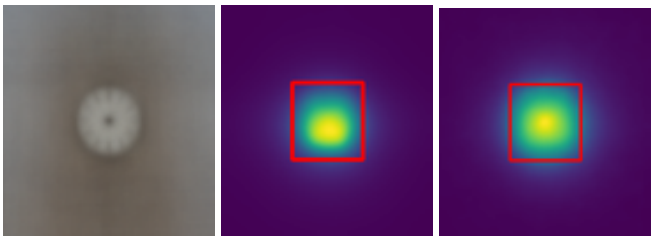
vase



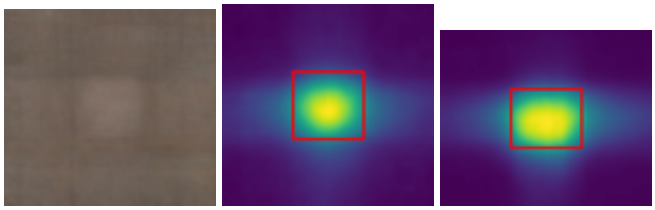
toaster



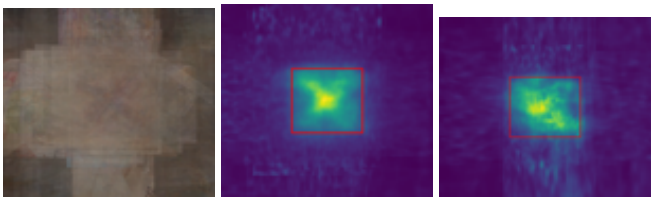
sink



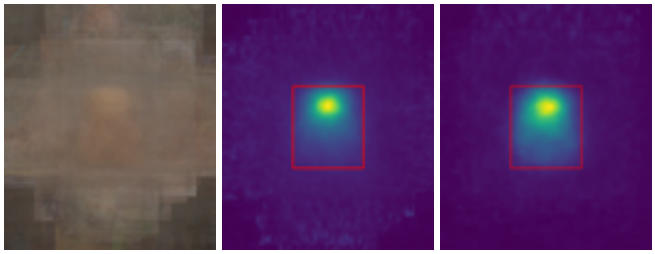
clock



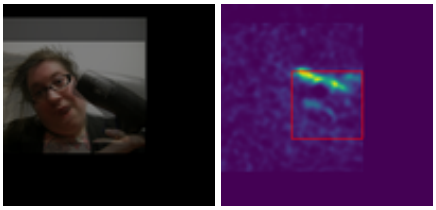
book



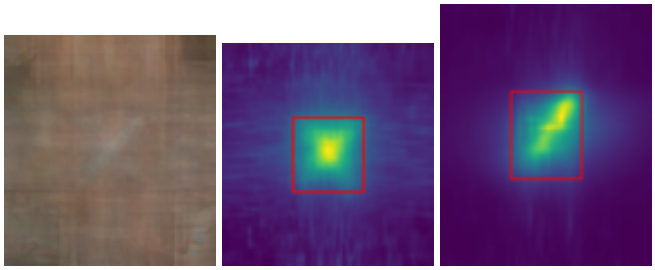
scissors



teddy bear



hair drier



toothbrush