Supplementary Material: Unconstrained Salient Object Detection via Proposal Subset Optimization

1. Proof of the Submodularity

According to Eqns. 10-12 in our paper, the objective function of the proposed optimization formulation can be represented as:

$$h(\mathbf{O}) = \sum_{i=1}^{n} \max_{x_i \in \widetilde{\mathbf{O}} \cup \{0\}} w_i(x_i) - \phi|\mathbf{O}| - \frac{\gamma}{2} \sum_{i,j \in \widetilde{\mathbf{O}}: i \neq j} \mathcal{K}_{ij}, \tag{1}$$

where $w_i(j) \triangleq \log P(x_i = j|I)$ and $\mathcal{K}_{ij}$ is shorthand for the bounding box similarity measure $\mathcal{K}(B_i, B_j)$. $\widetilde{\mathbf{O}}$ denotes the index set corresponding to the selected windows in $\mathbf{O}$.

Proposition 1. *$h(\mathbf{O})$ is a submodular function.*

Proof. Let

$$h(\mathbf{O}) = \sum_{i=1}^{n} \mathcal{A}_i(\mathbf{O}) + \phi\mathcal{B}(\mathbf{O}) + \gamma\mathcal{C}(\mathbf{O}), \tag{2}$$

where

$$\begin{aligned} \mathcal{A}_i(\mathbf{O}) &= \max_{x_i \in \widetilde{\mathbf{O}} \cup \{0\}} w_i(x_i), \\ \mathcal{B}(\mathbf{O}) &= -|\mathbf{O}|, \\ \mathcal{C}(\mathbf{O}) &= -\frac{1}{2} \sum_{i,j \in \widetilde{\mathbf{O}}: i \neq j} \mathcal{K}_{ij}. \end{aligned}$$

Because ϕ and γ are non-negative, it suffices to show $\mathcal{A}_i(\mathbf{O})$, $\mathcal{B}(\mathbf{O})$ and $\mathcal{C}(\mathbf{O})$ are all submodular, since the class of submodular functions is closed under non-negative linear combinations.

Recall that $\mathbf{O} \subseteq \mathbf{B} = \{B_i\}_1^n$, where $\mathbf{B}$ is the overall window proposal set. Let X and Y denote two subsets of $\mathbf{B}$, and $X \subseteq Y$. Also, let x denote an arbitrary window proposal such that $x \in \mathbf{B} \setminus Y$.

To show a function f is submodular, we just need to prove that $f(X \cup \{x\}) - f(X) \geq f(Y \cup \{x\}) - f(Y)$ [6, p. 766].

First, $\mathcal{B}(\mathbf{O})$ is submodular because

$$\begin{aligned} \mathcal{B}(X \cup \{x\}) - \mathcal{B}(X) &= -|X \cup \{x\}| + |X| \\ &= -|Y \cup \{x\}| + |Y| \\ &= \mathcal{B}(Y \cup \{x\}) - \mathcal{B}(Y). \end{aligned}$$

Second, $\mathcal{C}(\mathbf{O})$ is submodular because

$$\begin{aligned} \mathcal{C}(X \cup \{x\}) - \mathcal{C}(X) &= -\sum_{i \in \widetilde{X}} \mathcal{K}(B_i, B_{\widetilde{x}}) \\ &\geq -\sum_{i \in \widetilde{Y}} \mathcal{K}(B_i, B_{\widetilde{x}}) \\ &\geq \mathcal{C}(Y \cup \{x\}) - \mathcal{C}(Y), \end{aligned}$$

where $\widetilde{X}$, $\widetilde{Y}$ and $\widetilde{x}$ are the corresponding indices of X, Y and x w.r.t. $\mathbf{B}$. Note that $\mathcal{K}(B_i, B_{\widetilde{x}})$ is a similarity measure, and it is non-negative.

Lastly, we show that $\mathcal{A}_i(\mathbf{O})$ is submodular. Note that $\mathcal{A}_i$ is a monotone set function, so $\mathcal{A}_i(Y) \geq \mathcal{A}_i(X)$. Furthermore, $\mathcal{A}_i(X \cup \{x\}) = \max\{\mathcal{A}_i(X), \mathcal{A}_i^x\}$, where $\mathcal{A}_i^x \triangleq \mathcal{A}_i(\{x\})$. Thus,

$$\begin{aligned} &\mathcal{A}_i(Y \cup \{x\}) - \mathcal{A}_i(X \cup \{x\}) \\ =& \max\{\mathcal{A}_i(Y), \mathcal{A}_i^x\} - \max\{\mathcal{A}_i(X), \mathcal{A}_i^x\} \\ \leq& \mathcal{A}_i(Y) - \mathcal{A}_i(X). \end{aligned}$$

It is easy to see the last inequality by checking the cases when $\mathcal{A}_i^x \leq \mathcal{A}_i(X)$, $\mathcal{A}_i(X) < \mathcal{A}_i^x \leq \mathcal{A}_i(Y)$ and $\mathcal{A}_i(Y) < \mathcal{A}_i^x$ respectively. Then it follows that

$$\mathcal{A}_i(X \cup \{x\}) - \mathcal{A}_i(X) \geq \mathcal{A}_i(Y \cup \{x\}) - \mathcal{A}_i(Y).$$

Therefore, A_i is submodular. □

2. Merging Annotated Bounding Boxes

The MSRA [5] and DUT-O [9] datasets provide raw bounding box annotations from different subjects. To obtain a set of ground truth windows for each image, we use a greedy algorithm to merge bounding box annotations labeled by different subjects.

Let $\mathbf{B} = \{B_i\}_i^n$ denote the bounding box annotations for an image. For each bounding box B_i, we calculate an overlap score:

$$S_i = \sum_{j:j \neq i} \text{IOU}(B_i, B_j).$$

Based on the overlap score, we do a greedy non-maximum-suppression with the IOU threshold of 0.5 to get a set

of candidate windows. To suppress outlier annotations, a candidate window B_i is removed if there are fewer than two other windows in $\mathbf{B}$ that significantly overlap with B_i (IOU > 0.5). The remaining candidates are output as the ground truth windows for the given image.

3. CNN Model Training Details

We generate 100 exemplar windows by doing K-means clustering on the bounding box annotations of the SOS training set. The training images are resized to 224×224 regardless of their original dimensions. Training images are augmented by flipping and random cropping. Bounding box annotations that overlap with the cropping window by less than 50% are discarded. We use Caffe [3] to train the CNN model with a minibatch size of 8 and a fixed base learning rate of 10^{-4}. We fine-tune all the fully-connected layers together with conv5_1, conv5_2 and conv5_3 layers by backpropagation. Other training settings are the same as in [7]. We fine-tune the model on the ILSVRC-2014 detection dataset for 230K iterations, when the validation error plateaus. Then we continue to fine-tune the model on the SOS training dataset for 2000 iterations, where the iteration number is chosen via 5-fold cross validation. Fine-tuning takes about 20 hours on the ILSVRC dataset, and 20 mins on the SOS dataset using a single NVIDIA K40C GPU.

4. The Maximum Marginal Relevance Baseline

In our experiments, the Maximum Marginal Relevance (MMR) baseline follows the formulation in [1]. The MMR re-scores each proposal by iteratively selecting the proposal with maximum marginal relevance *w.r.t.* the previously selected proposals. The maximum marginal relevance is formulated by

$$\text{MMR} = \underset{h_i \in H \setminus H_p}{\arg\max} \left[s(h_i) - \theta \cdot \max_{h_j \in H_p} \text{IOU}(h_i, h_j) \right], \quad (3)$$

where H_p is the previously selected proposals. We optimize the parameter θ for the MMR baseline *w.r.t.* the AP score. For SalCNN, we use $\theta = 1.3$, and for MBox, we use $\theta = 0.05$.

5. Sample Detection Results

We show sample detection results of our method on the four datasets: MSO [10] (Figs. 1 and 2), VOC07 [2] (Figs. 3 and 4), DUT-O [9] (Figs. 5 and 6) and MSRA [5] (Figs. 7 and 8).

6. Results on PASCAL-S

For completeness, we further evaluate our method on the PASCAL-S dataset [4]. The images of PASCAL-S are from the PASCAL VOC07 dataset [2], and the ground truth is labeled based on the eye fixation data and the object segmentation masks provided by VOC07. This dataset contains some images with multiple objects, and most of its images contain at least one salient object.

Table 1: AP scores on the PASCAL-S dataset

MBox+NMS	MBox+MAP	SalCNN+NMS	SalCNN+MAP
.605	**.624**	.547	.599

Table 1 show the results of our MAP formulation compared with the NMS baseline for our SalCNN and MBox [8]. We find that our MAP formulation consistently improves over the NMS baseline for both MBox and our SalCNN on this dataset. MBox+MAP is slightly better than SalCNN+MAP on this dataset. Note that the ground truth annotations of PASCAL-S are still limited to the 20 categories in PASCAL VOC that MBox is trained on and optimized for.

References

[1] J. Carreira and C. Sminchisescu. Cpmc: Automatic object segmentation using constrained parametric min-cuts. *PAMI*, 34(7):1312–1328, 2012.

[2] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results. http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html.

[3] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014.

[4] Y. Li, X. Hou, C. Koch, J. M. Rehg, and A. L. Yuille. The secrets of salient object segmentation. In *CVPR*, 2014.

[5] T. Liu, Z. Yuan, J. Sun, J. Wang, N. Zheng, X. Tang, and H.-Y. Shum. Learning to detect a salient object. *PAMI*, 33(2):353–367, 2011.

[6] A. Schrijver. *Combinatorial optimization: polyhedra and efficiency*, volume 24. Springer Science & Business Media, 2003.

[7] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015.

[8] C. Szegedy, S. Reed, D. Erhan, and D. Anguelov. Scalable, high-quality object detection. *arXiv preprint arXiv:1412.1441*, 2014.

[9] C. Yang, L. Zhang, H. Lu, X. Ruan, and M.-H. Yang. Saliency detection via graph-based manifold ranking. In *CVPR*, 2013.

[10] J. Zhang, S. Ma, M. Sameki, S. Sclaroff, M. Betke, Z. Lin, X. Shen, B. Price, and R. Měch. Salient object subitizing. In *CVPR*, 2015.

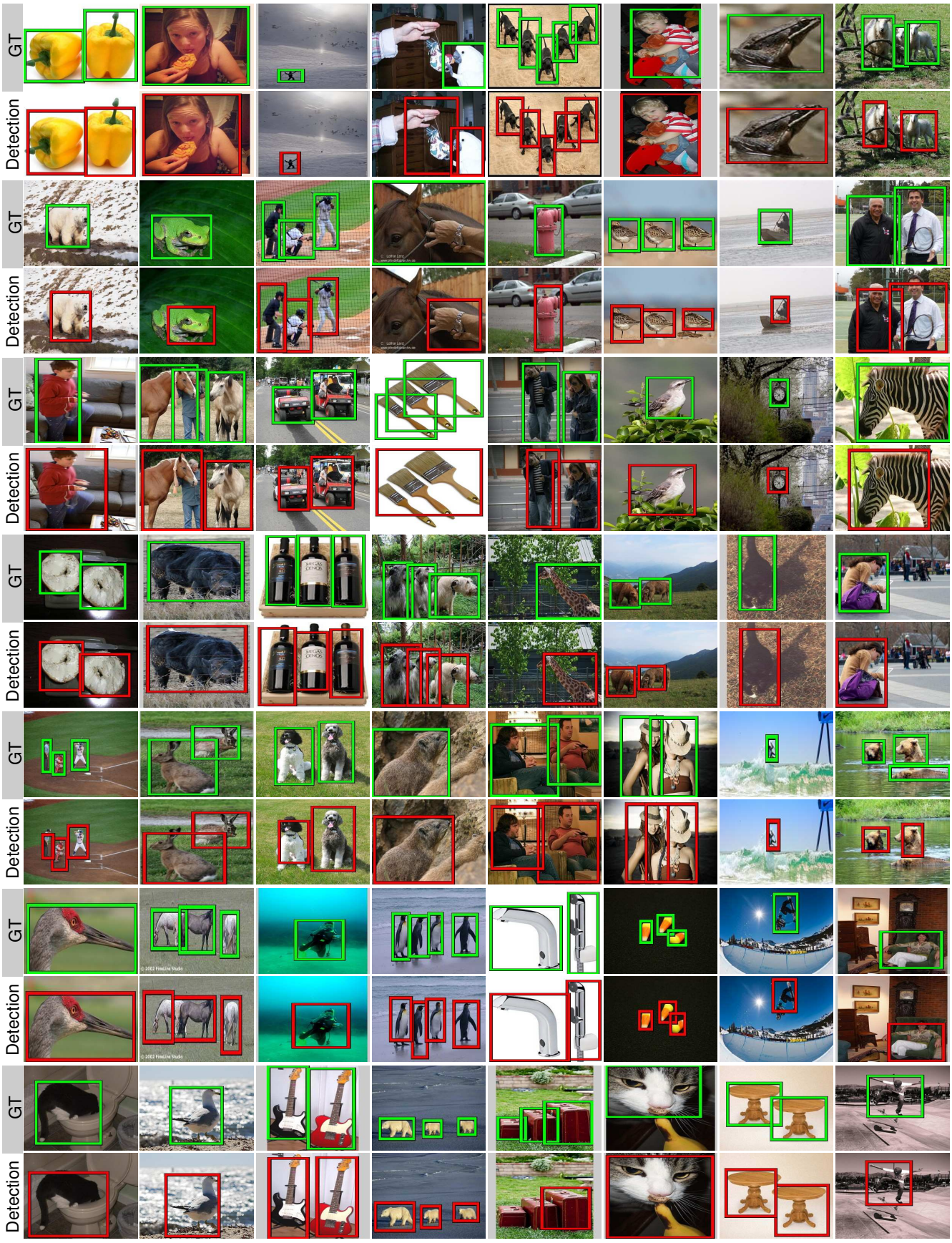

Figure 1: Sample detection results of our method when $\lambda = 0.1$ on the MSO dataset [10].

Figure 2: Sample detection results of our method when $\lambda = 0.1$ on the MSO dataset [10].

Figure 3: Sample detection results of our method when $\lambda = 0.1$ on the VOC07 dataset [2].

Figure 4: Sample detection results of our method when $\lambda = 0.1$ on the VOC07 dataset [2].

Figure 5: Sample detection results of our method when $\lambda = 0.1$ on the DUT-O dataset [9].

Figure 6: Sample detection results of our method when $\lambda = 0.1$ on the DUT-O dataset [9].

Figure 7: Sample detection results of our method when $\lambda = 0.1$ on the MSRA dataset [5].

Figure 8: Sample detection results of our method when $\lambda = 0.1$ on the MSRA dataset [5].