Supplement for Learning optimal seeds for diffusion-based salient object detection

Song Lu
SVCL Lab, UCSD
sol050@ucsd.edu

Vijay Mahadevan
Yahoo Labs
vmahadev@yahoo-inc.com

Nuno Vasconcelos
SVCL Lab, UCSD
nuno@ucsd.edu

1. Algorithm

In this section, we discuss the details of the learning algorithm of Section 3.3. The optimal weight vector minimizes the large-margin structured cost

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} E(\mathbf{w}) \tag{1}$$

where

$$E(\mathbf{w}) = \frac{\mathbf{1}}{\mathbf{2}}\alpha \parallel \mathbf{w} \parallel^{\mathbf{2}} + \sum_k \sum_{\{ij|\delta_i^k=1,\delta_j^k=0\}} \lfloor 1 - (y_i(\mathbf{x}^{(k)}) - y_j(\mathbf{x}^{(k)}) \rfloor_+, \tag{2}$$

$\lfloor x \rfloor_+ = \max(0, x)$ and

$$\mathbf{y}(\mathbf{x}) = \mathbf{A}(\mathbf{x})\mathbf{F}(\mathbf{x})\mathbf{w} \tag{3}$$

This can be written as

$$E(\mathbf{w}) = \frac{\mathbf{1}}{\mathbf{2}}\alpha \parallel \mathbf{w} \parallel^{\mathbf{2}} + \sum_k \sum_{\{ij|\delta_i^k=1,\delta_j^k=0\}} \lfloor 1 - (\xi_i^{(k)} - \xi_j^{(k)})\mathbf{F}^{(k)}\mathbf{w} \rfloor_+ \tag{4}$$

$$= \frac{1}{2}\alpha \parallel \mathbf{w} \parallel^2 + \sum_k \sum_{(i,j)\in S_k} (1 - (\xi_i^{(k)} - \xi_j^{(k)})\mathbf{F}^{(k)}\mathbf{w}) \tag{5}$$

where $\xi_i^{(k)} = (\mathbf{A}_i(\mathbf{x}^{(k)})$ is the i^{th} row of the propagation matrix for the k^{th} image, $\mathbf{F}^{(k)}$ the matrix of feature responses to the image, and $S_k = \{(i,j)|\delta_i^k = 1, \delta_j^k = 0, (\xi_i^{(k)} - \xi_j^{(k)})\mathbf{F}^{(k)}\mathbf{w} < 1\}$.

The optimization is performed by gradient descent. The gradient of (5) with respect to $\mathbf{w}$ is

$$\nabla_{\mathbf{w}} E(\mathbf{w}) = \alpha\mathbf{w} + \sum_{\mathbf{k}} \sum_{(\mathbf{i},\mathbf{j})\in\mathbf{S_k}} \left((\xi_{\mathbf{j}}^{(\mathbf{k})} - \xi_{\mathbf{i}}^{(\mathbf{k})})\mathbf{F}^{(\mathbf{k})}\right)^{\mathbf{T}},$$

and the gradient update equation

$$\mathbf{w}^{t+1} = \mathbf{w}^t - \lambda_0 * \nabla_{\mathbf{w}} E(\mathbf{w}) \tag{6}$$

Algorithm 1 learning weight for different features

Input: Propagation matrices $\{\mathbf{A}(\mathbf{x}^{(1)}), \ldots, \mathbf{A}(\mathbf{x}^{(n)})\}$, feature matrices $\{\mathbf{F}(\mathbf{x}^{(1)}), \ldots, \mathbf{F}(\mathbf{x}^{(n)})\}$, α.
Initialization:
Set $S_{diff} = 1, S = -1, \mathbf{w} = \mathbf{0}, \lambda_0 = 0.00003$
Iteration:
while $S_{diff} > 0.0001$ **do**
  $s = \frac{1}{2}\alpha||\mathbf{w}||_2$
  $\nabla_{\mathbf{w}} = 0$
  for all $k \in \{1, \ldots, n\}$ **do**
    $k_0 = 0, \gamma = \mathbf{0}$
    for all $i \in \{i|\delta_i^k = 1\}, j \in \{j|\delta_j^k = 1\}$ **do**
      $d = (\xi_i^{(k)} - \xi_j^{(k)})\mathbf{F}^{(k)}\mathbf{w}$
      if $d < 1$ **then**
        $k_0 \leftarrow k_0 + 1$
        $\gamma = \gamma + \left((\xi_j^{(k)} - \xi_i^{(k)})\mathbf{F}^{(k)}\right)^T$
        $s = s + 1 - d$
      end if
    end for
    $\nabla_{\mathbf{w}} = \nabla_{\mathbf{w}} + \frac{1}{k_0}\gamma$
  end for
  $\nabla_{\mathbf{w}} = \nabla_{\mathbf{w}} + \alpha\mathbf{w}$
  $\mathbf{w} = \mathbf{w} - \lambda_0\nabla_{\mathbf{w}}$
  $S_{diff} = |s - S|$
  $S = s$
end while
Output: $\mathbf{w}$

where λ_0 is a learning rate. Convergence is declared when the cost of (5) decreases by less than a threshold in consecutive iterations. Algorithm 1 summarizes the learning procedure.

2. More experiments

In this section, we present results of a number of experiments that analyze in greater detail the role of the different features in the performance of the proposed algorithm.

Table 1. Feature list

feature description	feature No.
contrast to neighboring superpixels	1-25
contrast to left boundary	26-50
contrast to right boundary	51-75
contrast to top boundary	76-100
contrast to bottom boundary	101-125
geometry features	126-160
Element distribution	161-163
Element uniqueness	164-166
Pattern distinctness (2 scales)	167-168
Color distinctness (1 scale)	169
Center bias	170-172
Backgroundness	173-177
eye fixation feature	178

Table 2. Contrast features

feature description	feature No.
difference of average RGB value	1-3
difference of average RGB value	4-6
Chi distance of LAB histogram	7
Chi distance of hue histogram	8
Chi distance of saturation histogram	9
difference of average texton	10-24
Chi distance of text on histogram	25

2.1. Features

Table 1 summarizes the features used in our implementation. Features 1 to 125 are five sets of contrast features. These measure the contrast between the visual content of a superpixel and a set of other superpixels. The latter can be one of five sets: the neighboring superpixels or the superpixels on the four image boundaries. In each case, a set of 25 features are computed. These are described in Table 2. They involve measures of color and texture contrast, typically a χ^2 difference between histograms of color or texton response. Given two histograms $\mathbf{h}^a$ and $\mathbf{h}^b$, this is defined as

$$\chi^2(\mathbf{a}, \mathbf{b}) = \frac{1}{2} \sum_{m=1}^{K} \frac{[h^a(m) - h^b(m)]^2}{h^a(m) + h^b(m)} \qquad (7)$$

Features 126 to 160 capture geometric properties of superpixels. These features were proposed in the regional property descriptor of [5]). They are listed in Table 3. The geometric properties accounted for include various areas, aspect ratio, and descriptors of the spatial coordinates of superpixels, as well some measures of the spatial distribution of colors and texture. Features 161 to 177 are the mid-level vision features discussed in Section 3.5. of the paper. Finally, the eye fixation feature is the bottom-up saliency map of Section 3.4.

The importance of the various features was analyzed through a number of experiments. The saliency detector was implemented using the saliency seeds produced by each feature alone. The features were then ranked by the resulting AUC score on the VOC2008_1023 dataset. Figure 1 shows the ranked scores for the 15 best performing features. The top feature was the bottom-up saliency map, confirming the expectation that predicted eye fixations are informative of object saliency. Interestingly, the majority of the remaining features in the top 15 list were measures of contrast between image superpixels and superpixels on the image boundaries. This is indicative of the fact that, because photographers tend to capture salient objects in the center of the image, image boundaries tend to contain background. Note, however, that the center bias features commonly used to account for this effect in the saliency literature (170-172) do not appear in the top 15. Beyond bottom-

Table 3. Geometry features

feature description	feature No.
normalized area	126
normalized superpixel number	127
average normalized x coordinates	128
average normalized y coordinates	129
10^{th} percentile of normalized x coordinates	130
10^{th} percentile of normalized y coordinates	131
90^{th} percentile of normalized x coordinates	132
90^{th} percentile of normalized y coordinates	133
aspect ratio	134
normalized neighbor superpixel number	135
normalized area of neighbor region	136
variance of RGB	137-139
variance of Lab	140-142
variance of hue and saturation	143-144
variance of LM filter response	145-159
normalized area	160

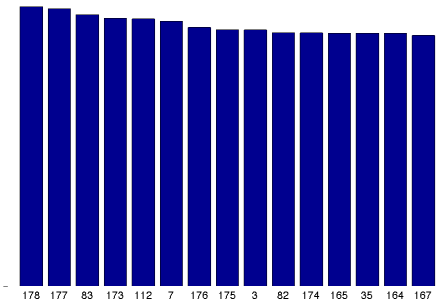

Figure 1. AUC score of the 15 best performing features on the VOC2008 dataset.

Table 4. Performance of different feature combination methods: AUC/AP

AUC/AP	MSRA5000	SOD	SED1	SED2	VOC2008_1023
MeanseedProp	0.8935/0.7718	0.7343/0.5721	0.8289/0.7445	0.7603/0.6127	0.7001/0.65464
SalseedProp	0.9058/0.8136	0.8175/0.6688	0.9176/0.8537	0.8806/0.7500	0.7908/0.6421
OptseedProp	**0.9615/0.8790**	**0.8684/0.7019**	**0.9530/0.8905**	**0.9058/0.8062**	**0.8181/0.6556**

up saliency and contrast to boundary, the remaining top 15 features are mostly measures of contrast to neighboring superpixels. This was expected, given the well known role of center-surround processes in saliency.

2.2. Learning

We next evaluated the impact of learning feature combinations in the performance of the object saliency detector. For this, we compared the performance of the detector based on the proposed learning algorithm with that of a detector without learning. This was implemented by simply assuming uniform weights, i.e. setting all the entries of $\mathbf{w}$ to $1/n$, where n is the number of features. In this case, the saliency seed image is simply the average of all the feature maps. The performance of the resulting detector (denoted *MeanseedProp*) is compared to that of the optimal detector (*OptseedProp*), for all the datasets discussed in Section 4 of the paper, in Table 4. Also shown is the performance obtained with the best feature only, i.e. using the bottom-up saliency map to determine the saliency seeds (denoted *SalseedProp*). Note that simply using many features provides no guarantee of good performance. In fact, MeanseedProp performs worse than using only the eye fixation predictions to determine saliency seeds. On the other hand, the proposed learning algorithm learns an effective seed map, producing a saliency detector with substantially higher AUC than those produced by the competing feature combination mechanisms. Finally, Figure 2 presents a ranking of the top 15 features according to the learning algorithm, i.e. according to the weight w_i learned for each feature. The list of top features includes several of the features of Figure 1, i.e. the features of highest individual AUC score. Note that, as in Figure 1, the top three features are the bottom-up saliency map and two measures of contrast to boundary pixels.

3. Saliency detection results

We finish by presenting an extended comparison of the saliency maps obtained in different datasets. Figure 3 and 4 present a comparison between the proposed saliency detector and eight of the best performing methods in the literature - Gof [3], CB [4], HC [2], RC [2], GBMR [8], PCA [6], FT [1], and SF [7] - on the MSRA5000 dataset. Figure 5 and Figure 6 present a similar comparison on VOC2008_1023.

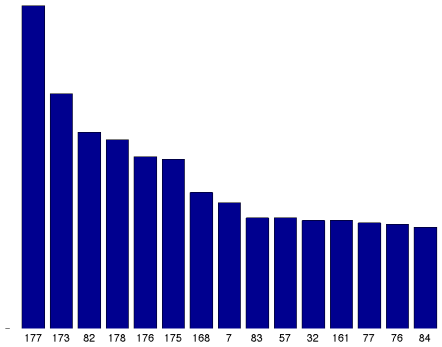

Figure 2. Weights assigned by the learning algorithm to the 15 features of largest weight.

References

[1] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk. Frequency-tuned salient region detection. In *CVPR*, pages 1597–1604. IEEE, 2009. 3

[2] M.-M. Cheng, G.-X. Zhang, N. J. Mitra, X. Huang, and S.-M. Hu. Global contrast based salient region detection. In *CVPR*, pages 409–416. IEEE, 2011. 3

[3] S. Goferman, L. Zelnik-Manor, and A. Tal. Context-aware saliency detection. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(10):1915–1926, 2012. 3

[4] H. Jiang, J. Wang, Z. Yuan, T. Liu, N. Zheng, and S. Li. Automatic salient object segmentation based on context and shape prior. In *BMVC*, volume 3, page 7, 2011. 3

[5] H. Jiang, J. Wang, Z. Yuan, Y. Wu, N. Zheng, and S. Li. Salient object detection: A discriminative regional feature integration approach. In *CVPR*, 2013. 2

[6] R. Margolin, A. Tal, and L. Zelnik-Manor. What makes a patch distinct? In *CVPR*, 2013. 3

[7] F. Perazzi, P. Krahenbuhl, Y. Pritch, and A. Hornung. Saliency filters: Contrast based filtering for salient region detection. In *CVPR*, pages 733–740. IEEE, 2012. 3

[8] C. Yang, L. Zhang, H. Lu, X. Ruan, and M.-H. Yang. Saliency detection via graph-based manifold ranking. In *CVPR*, 2013. 3

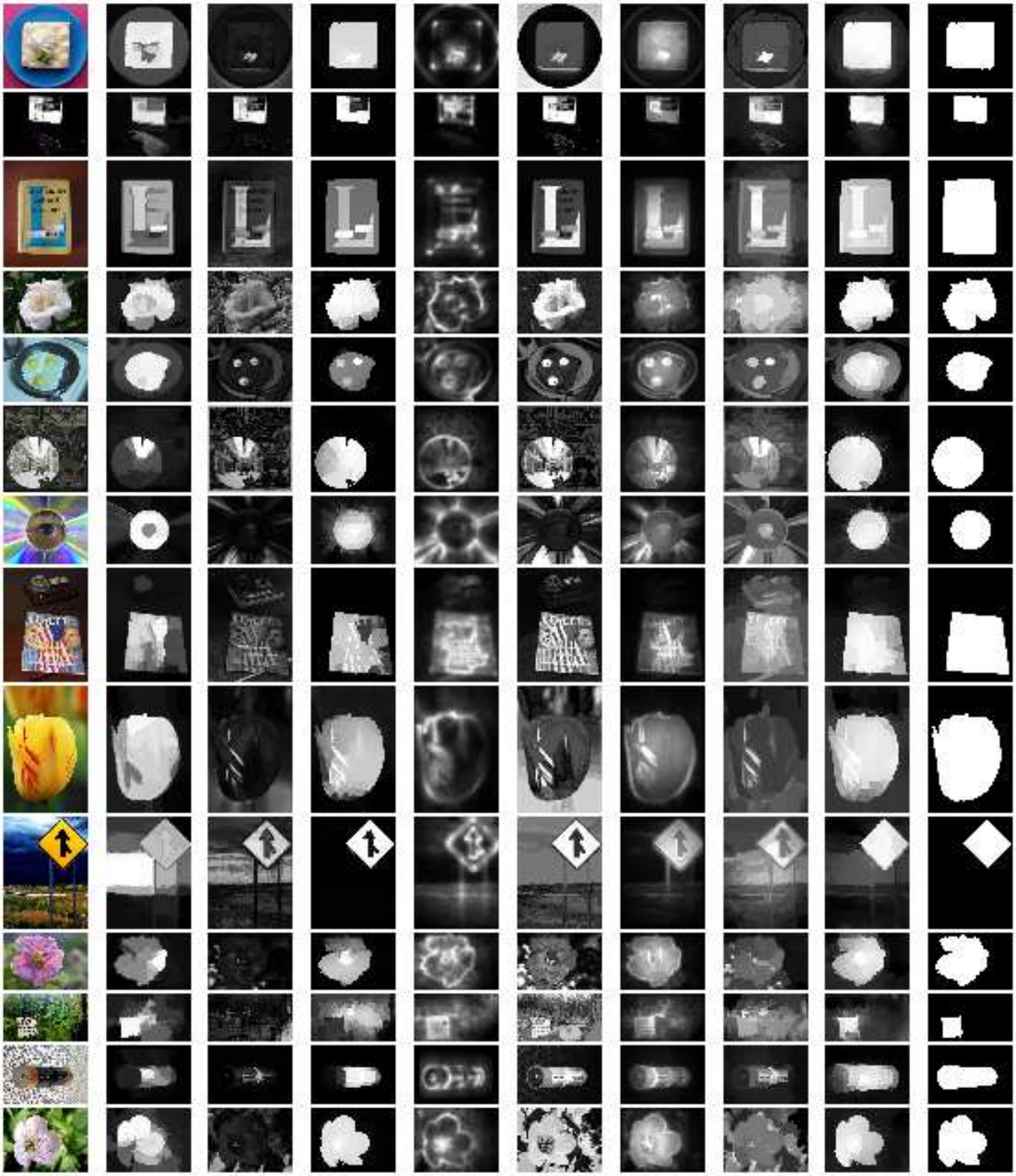

Figure 3. Comparison of results on representative images from MSRA5000 datasets . The original image is shown on the extreme left column. The other columns from left to right are the outputs of: 'CB','FT','GBMR','Gof','HC','PCA','RC','OptSeedProp (proposed)'. The binary ground truth is shown in the column on the far right.

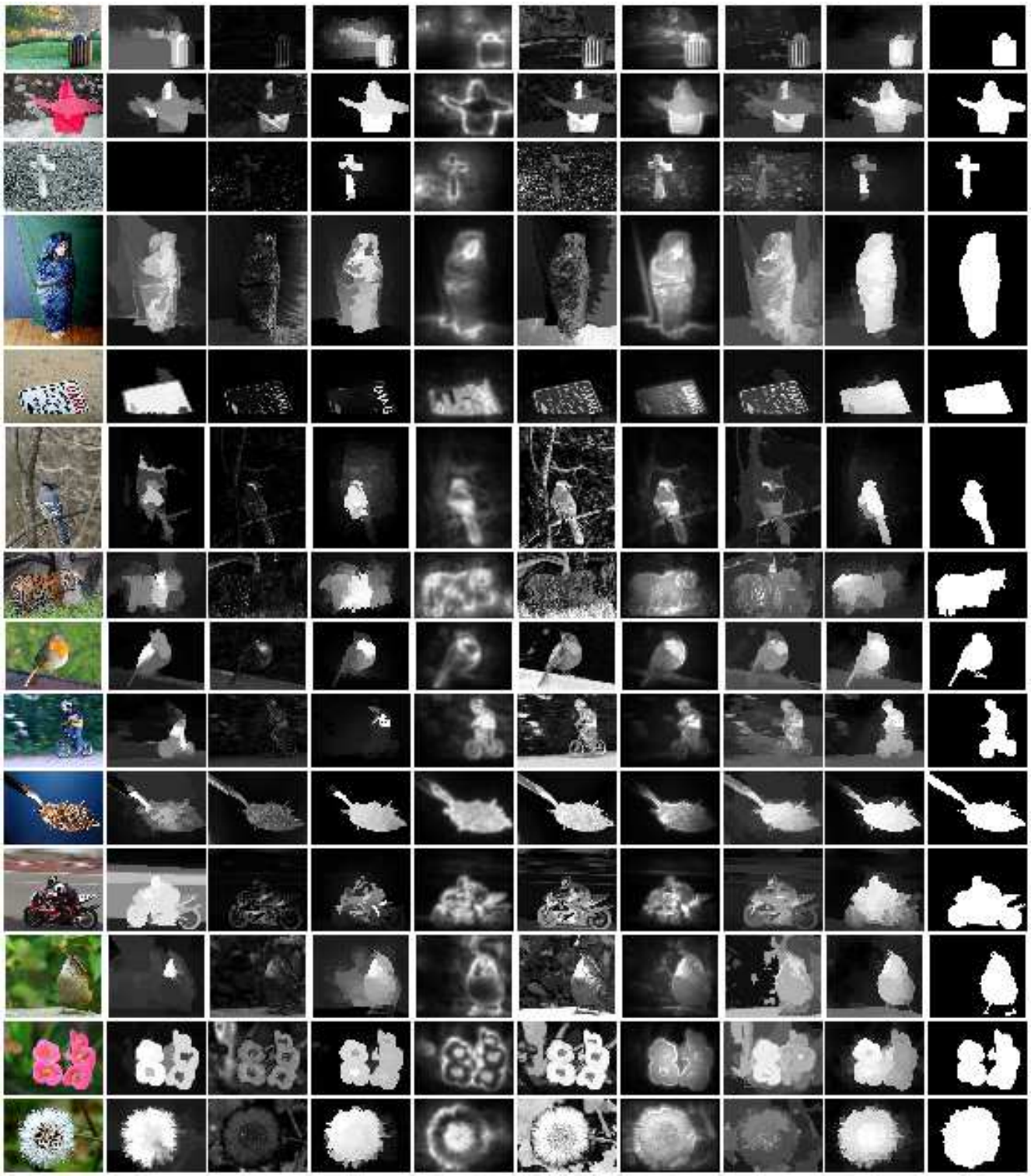

Figure 4. Comparison of results on representative images from MSRA5000 datasets . The original image is shown on the extreme left column. The other columns from left to right are the outputs of: 'CB','FT','GBMR','Gof','HC','PCA','RC','OptSeedProp (proposed)'. The binary ground truth is shown in the column on the far right.

Figure 5. Comparison of results on representative images from VOC2008_1023 datasets . The original image is shown on the extreme left column. The other columns from left to right are the outputs of: 'CB','FT','GBMR','Gof','HC','PCA','RC','OptSeedProp (proposed)'. The binary ground truth is shown in the column on the far right.

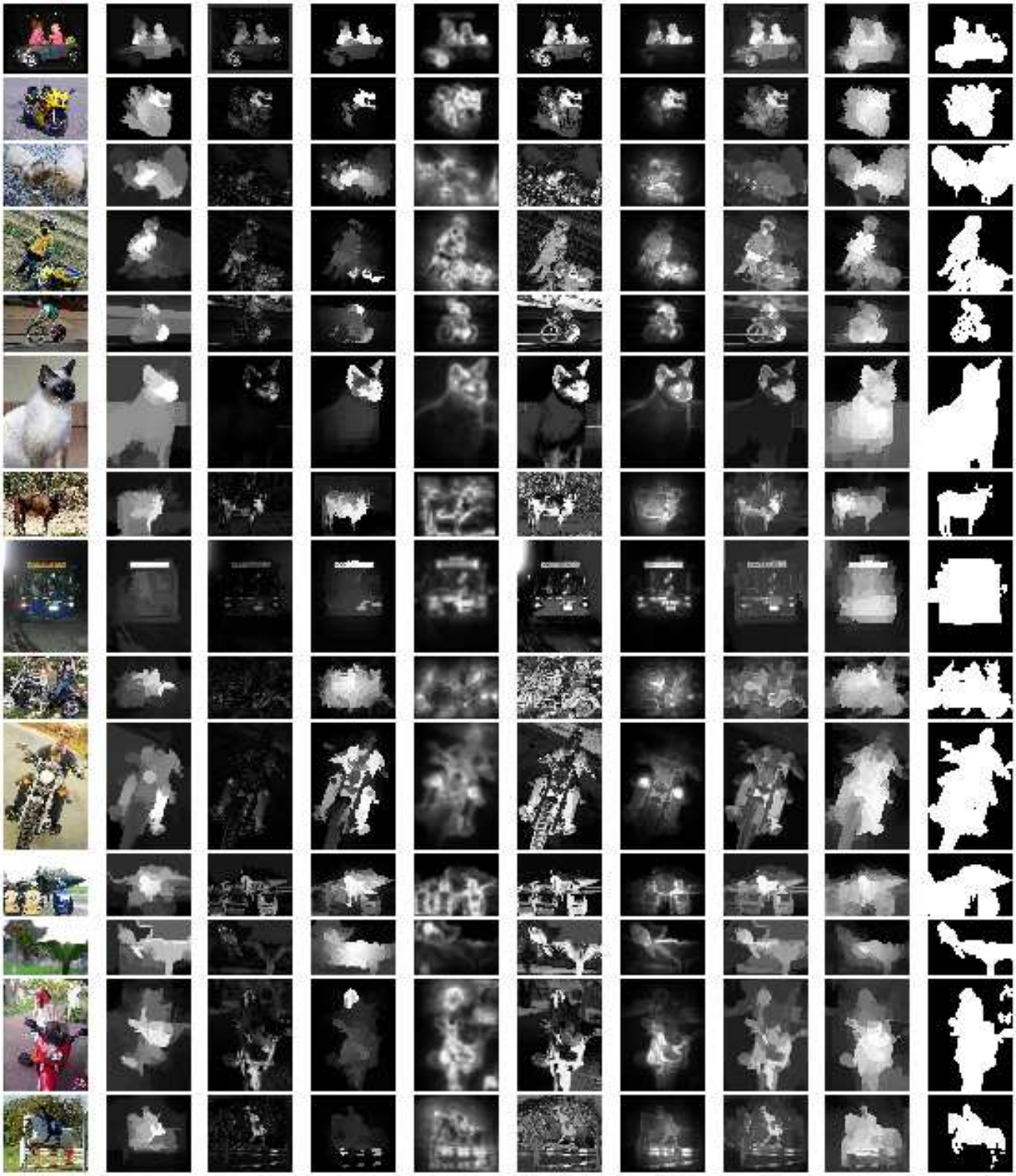

Figure 6. Comparison of results on representative images from VOC2008_1023 datasets . The original image is shown on the extreme left column. The other columns from left to right are the outputs of: 'CB','FT','GBMR','Gof','HC','PCA','RC','OptSeedProp (proposed)'. The binary ground truth is shown in the column on the far right.