Supplementary for: Encoding based Saliency Detection for Videos and Images

Thomas Mauthner, Horst Possegger, Georg Waltner, Horst Bischof
Institute for Computer Graphics and Vision, Graz University of Technology
{mauthner,possegger,waltner,bischof}@icg.tugraz.at

1. Supplementary

In the following, we summarize the additional explanations, evaluations and visualizations. We start with a description of evaluation metrics applied within our paper in Section 1.1. Next we summarize our experimental results on the Weizmann [2] data set, comparing with results recently proposed by [8]. Finally, we discuss in detail the additional experiment for cropping centered objects in the ASD dataset [1], and give visual examples and comparisons with competing approaches.

1.1. Evaluation Metrics

Within our experimental section several metrics are applied to evaluate saliency detection methods against each other. As ground truth annotations are given in different formats (*i.e.* coarse bounding boxes, detailed binary segmentation or eye-fixation maps), we apply the following metrics correspondingly. If ground truth segmentation is available, we compute precision/recall values as well as the area under curve (AUC) by varying thresholds to obtain binarized saliency maps and measure the overlap with the ground truth segmentation. For experiments where solely bounding box annotations are available, we add spanning bounding boxes to the binarized saliency map before computing the scores (denoted AUC-box, see Figure 1). For given eye-gaze ground truth data, we measure the exactness of the saliency maps by computing the normalized cross correlation (NCC).

AUC: For evaluation on segmented object ground truth, we compute true-positives (TP), false-positives (FP), true-negatives (TN) and false-negatives (FN) for each threshold image (see Figure 1, second row). The TP is the number of pixels with saliency values $\geq$ threshold and overlapping with the ground truth. In contrast, FN are all pixels within the ground truth region $<$ threshold. By varying this threshold one can compute recall-precision curves by

$$precision = \frac{TP}{TP + FP} \tag{1}$$

and

$$recall = \frac{TP}{TP + FN}. \tag{2}$$

The area under this recall precision curve is denoted as AUC. Although the saliency maps of different algorithms align nicely with the object region, the bounding box annotation causes many false-negatives, as depicted in Figure 1.

AUC-Box: Filling the binary saliency maps with spanning bounding boxes before computing TP, FP, TN and FN compensates for coarse annotation (see Figure 1). We denote this measure AUC-box within the paper. Given both scores for the UCF sports dataset, the reader may extract additional information about performance and robustness of methods.

NCC: If non-binary ground truth information is given, *e.g.* as eye-gaze tracking data, we apply normalized cross correlation for measuring performance. Eye-gaze data is generally given as a set of sparse local points of fixations or saccadic motions. As defined by the collectors of the data [4], we apply Gaussian blur for each gaze measurement to compensate measurement errors and create a smooth map.

1.2. Saliency for Activity Detection

We follow the recent evaluation of image and video saliency methods by Zhou *et al.* [8] on the Weizmann activity dataset [2] and compare our proposed encoding based saliency (EBS) method to the top-performing methods [3, 5, 8] of that study. The dataset contains videos of ten activities performed by nine actors captured with object-centered static cameras in front of a homogenous background. This simplifies the video saliency estimation to a foreground estimation problem. In fact, results in [8] have shown the superior performance of solely color-based methods, while video saliency approaches (which include motion information) perform worse. Figure 2 shows this bias of the evaluation strategy. Although our weighted saliency approach (EBSG) yields visually plausible results,

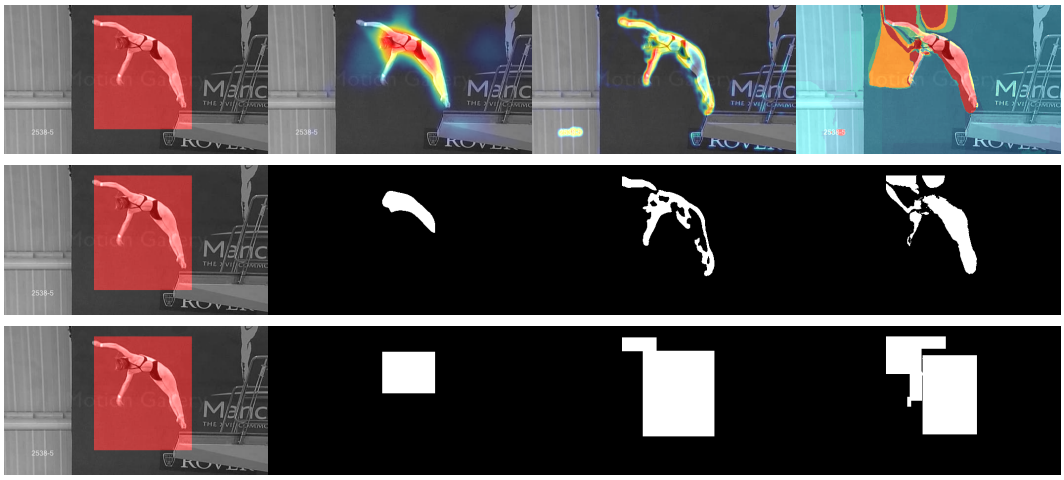


Figure 1: Comparison of AUC and AUC-box computation. Left column shows the ground truth bounding box annotation. Top row depicts results from individual saliency detection methods (f.l.t.r) our proposed EBSG, [6] and [5]. Second row shows true-positives for different thresholds. Especially non-segmentation results like column two and three have many false-negatives, hence low recall and AUC values. Applying spanning bounding-boxes around sub-segments compensates for the coarse segmentation, allowing for a fairer comparison (third row).

the segmentation ground truth prefers fully segmented objects. In particular, considering purely local activity (*e.g.* hand waving) the motion-based saliency focuses on such active regions, which results in reduced performance metrics on the binary ground truth masks as these cover the whole person. However, our solely color-based approach EBS(color) shows competitive results in comparison to the top-performing methods.

1.3. Salient Object Detection In Images

Figure 3 summarizes the results for different parameter combinations between color-space, bins per channel, number of encoding vectors and number of nearest neighbor encoding vectors applied for computing the saliency values of pixels. Although this evaluation creates over 50 different combinations, all results are within state-of-the-art which underlines the robustness of our proposed method. As discussed within our paper, salient object datasets are biased towards centered objects without connection to the image border. All methods performing favorable compared to our proposed EBS methods, exploit this circumstance. To evaluate robustness of methods if this assumption is violated, and to compare our EBSG against top performing BMS [7], we created two datasets by cropping images of the ASD dataset such that salient objects are located near the borders. Two cropping levels are tested: First, salient objects touch the closest image border and second, intersect the closest border by 5 pixels. Directly visible in Figure 4, and depicted in Figure 4c of our main paper, the robustness of BMS decreases drastically while EBSG stays almost constant within the first test and decreases slightly for severe *out of center* objects. Additional visual comparisons can be found in Figures 5 and 6.

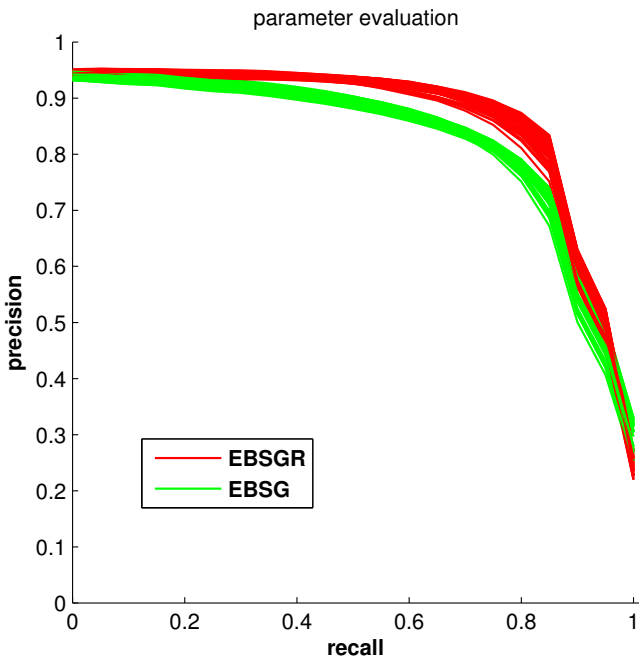
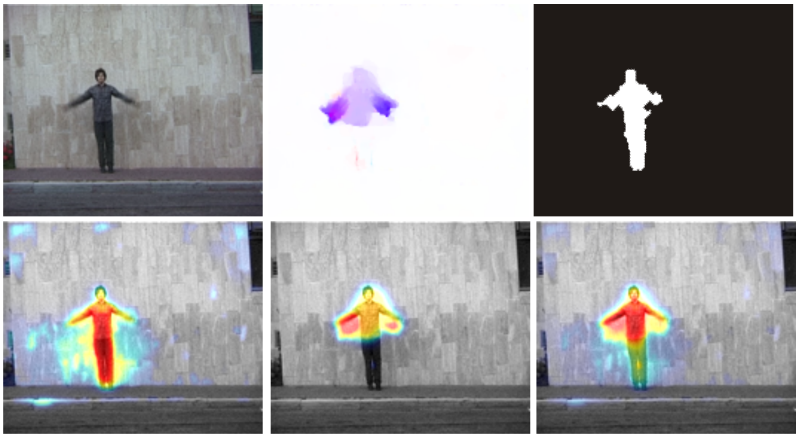
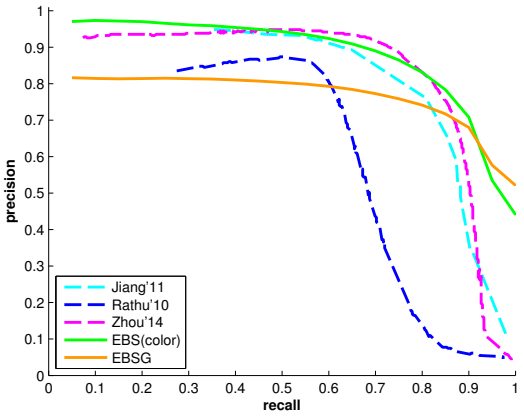


Figure 3: Results for different parameter configurations on the ASD data set.



(a)



(b)

Figure 2: Weizmann video saliency. Top row (a): image, motion, and ground truth segmentation. Bottom row: saliency results for taking color, motion or combining both cues by our proposed weighting scheme. Average recall precision curves are shown in (b). Our EBS method taking solely color information performs favorably. See text for further discussion.

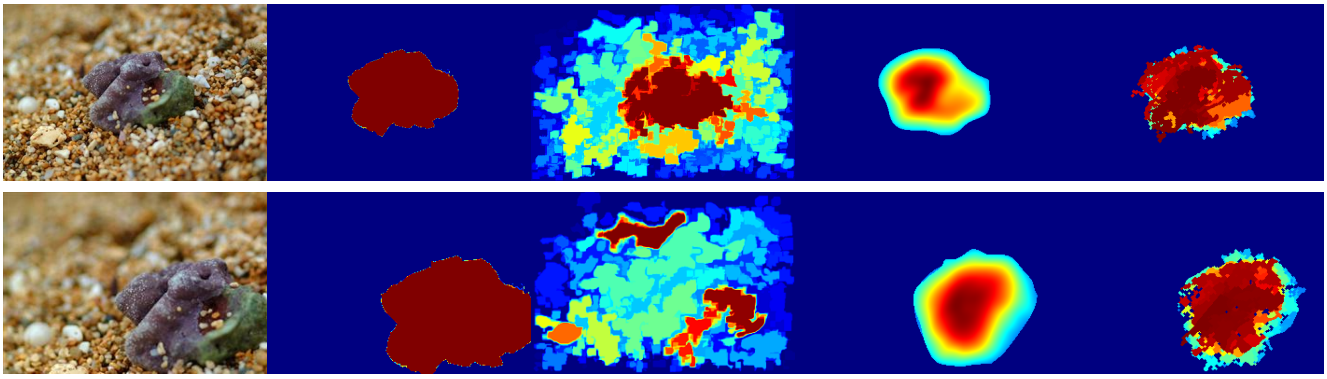


Figure 4: Visual comparison of EBS and BMS [7] for detecting salient objects which intersect with image border. Top row shows (f.l.t.r) original input image, ground truth, results for BMS and EBSG. Right-most column shows EBSGR results using encoding information for over-segmentation and propagating high saliency values within these segments. Bottom row shows cropped image with object attached to border. Still the object defines the visual salient part of the image, but performance of BMS strongly decreases while EBSG performs well. As we fill regions after computing the saliency with EBSGR, our segmentation results contains details of the object.

References

[1] R. Achanta, S. Hemami, F. Estrada, and S. Süsstrunk. Frequency-tuned Salient Region Detection. In *CVPR*, 2009. 1

[2] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri. Actions as Space-Time Shapes. *PAMI*, 29(12):2247–2253, 2007. 1

[3] H. Jiang, J. Wang, Z. Yuan, N. Zheng, and S. Li. Automatic Salient Object Segmentation Based on Context and Shape Prior. In *BMVC*, 2011. 1

[4] S. Mathe and C. Sminchisescu. Dynamic Eye Movement Dataset and Learnt Saliency Models for Visual Action Recognition. In *ECCV*, 2012. 1

[5] E. Rahtu, J. Kannala, M. Salo, and J. Heikkilä. Segmenting Salient Objects from Images and Videos. In *ECCV*, 2010. 1, 2

[6] W. Sultani and I. Saleemi. Human Action Recognition across Datasets by Foreground-weighted Histogram Decomposition. In *CVPR*, 2014. 2

[7] J. Zhang and S. Sclaroff. Saliency Detection: A Boolean Map Approach. In *ICCV*, 2013. 2, 3, 4, 5

[8] F. Zhou, S. B. Kang, and M. F. Cohen. Time-Mapping using Space-Time Saliency. In *CVPR*, 2014. 1

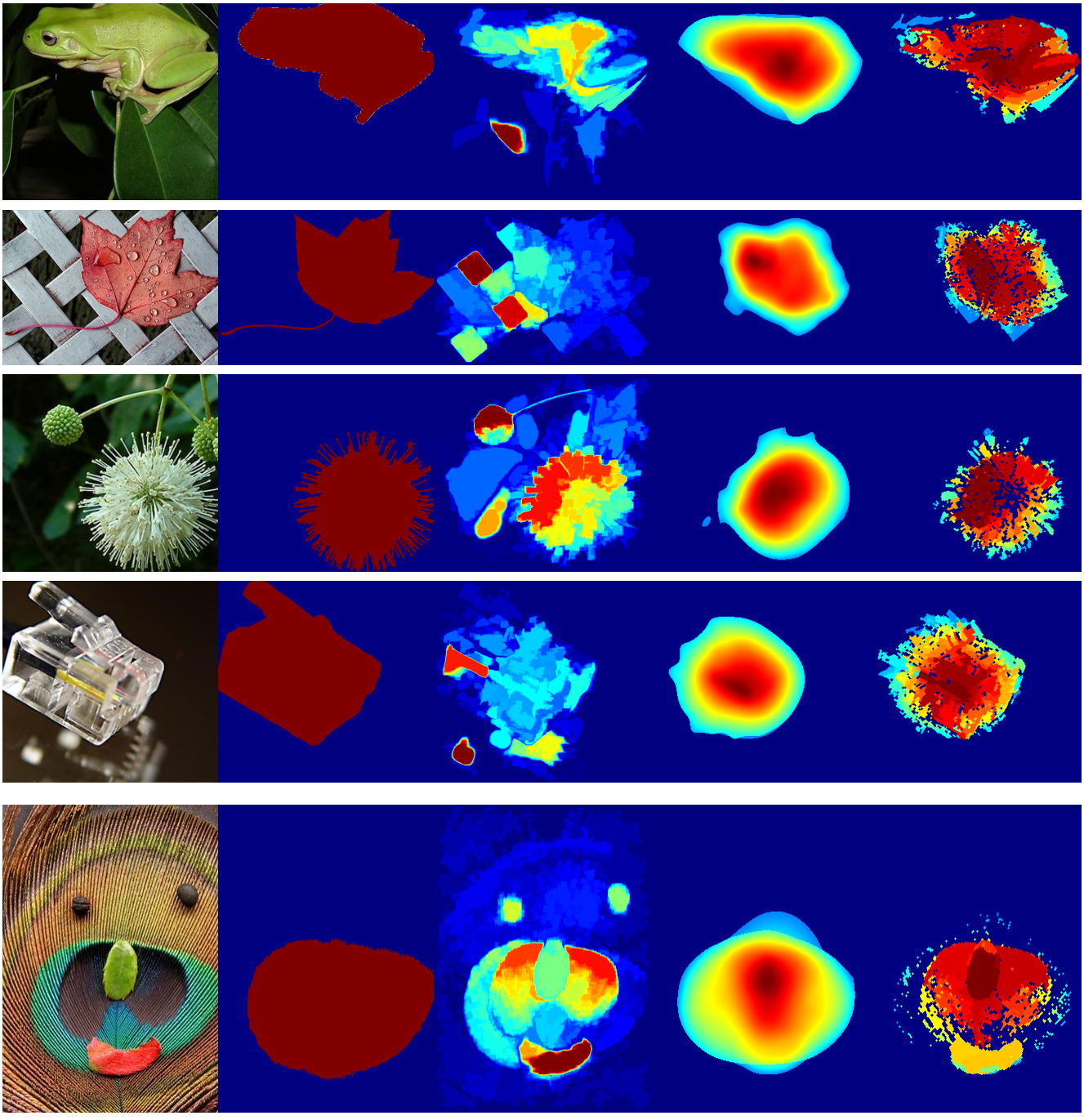


Figure 5: Visual comparison of EBS with BMS [7] on cropped images of complex textured examples (f.l.t.r): Input image, ground truth, BMS, EBSG. Final column shows EBSGR using encoding information for over-segmentation and propagating high saliency values within these segments.

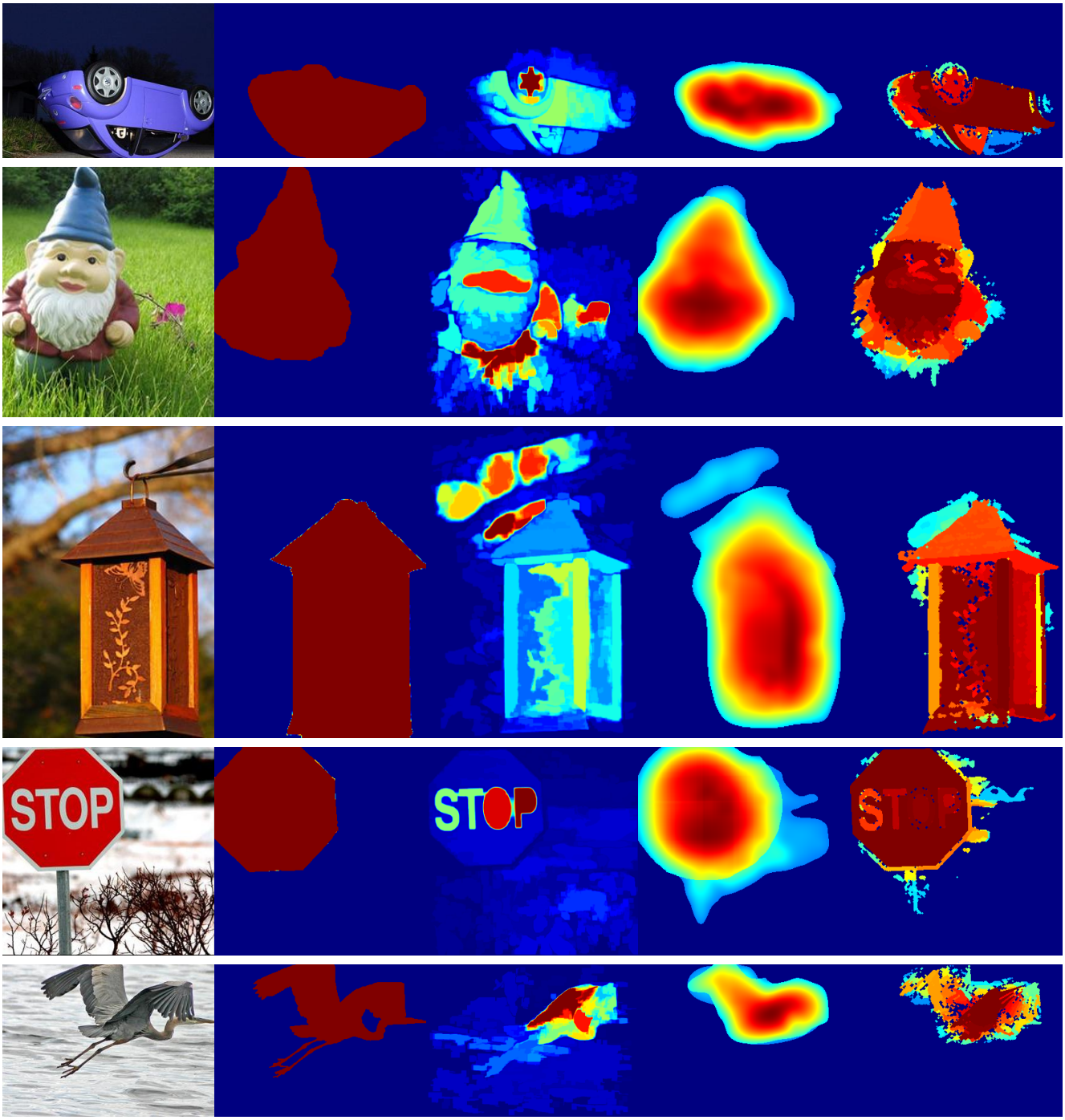


Figure 6: Further visual comparison of EBS with BMS [7] on cropped images, for detecting salient objects connected to the image border (f.l.t.r): Input image, ground truth, BMS, EBSG. Right-most column shows EBSGR using encoding information for over-segmentation and propagating high saliency values within these segments.