Weakly Supervised Saliency Detection with A Category-Driven Map Generator

Kuang-Jui Hsu[1,2]
https://www.citi.sinica.edu.tw/pages/kjhsu/
Yen-Yu Lin[1]
https://www.citi.sinica.edu.tw/pages/yylin/
Yung-Yu Chuang[2]
https://www.csie.ntu.edu.tw/~cyy/

[1] Academia Sinica, Taipei, Taiwan
[2] National Taiwan University, Taipei, Taiwan

Abstract

Top-down saliency detection aims to highlight the regions of a specific object category, and typically relies on pixel-wise annotated training data. In this paper, we address the high cost of collecting such training data by presenting a weakly supervised approach to object saliency detection, where only image-level labels, indicating the presence or absence of a target object in an image, are available. The proposed framework is composed of two deep modules, an image-level classifier and a pixel-level map generator. While the former distinguishes images with objects of interest from the rest, the latter is learned to generate saliency maps so that the training images masked by the maps can be better predicted by the former. In addition to the top-down guidance from class labels, the map generator is derived by also referring to other image information, including the background prior, area balance and spatial consensus. This information greatly regularizes the training process and reduces the risk of overfitting, especially when learning deep models with few training data. In the experiments, we show that our method gets superior results, and even outperforms many strongly supervised methods.

1 Introduction

Object saliency detection has been an active topic in the field of computer vision. The detected saliency maps highlight the regions of objects attracting people. They are crucial to various computer vision applications such as image retargeting [50], visual tracking [16], object segmentation [6, 44] and object recognition [35], since objects of interest are kept while irrelevant background is filtered out.

Object saliency detection methods can be roughly divided into the bottom-up and the top-down groups. The bottom-up methods rely on merely the information computed from images for detection. They seek object regions by finding their distinct characteristics from the background. Despite the generality, methods of this group often fail if the difference between objects and background is subtle. By contrast, top-down approaches [9, 10, 15, 25, 46] are category-aware. They utilize the prior knowledge about a target object category for saliency detection, and do not suffer from the aforementioned limitation. However, the top-down

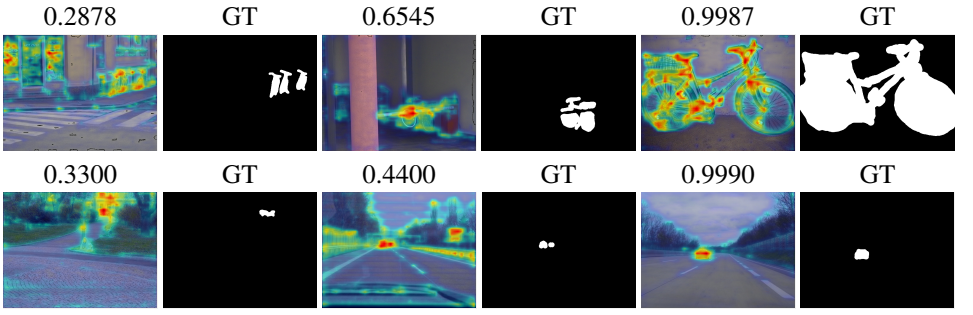


Figure 1: Three examples of the detected saliency maps and their ground truth for the category *bike* (top row) and *car* (bottom row). On the top of each map, we show the score by applying the category classifier to the image with its non-salient regions removed. The better the non-salient areas are removed, the higher the classification scores are.

methods need training data in the form of pixel-wise annotations, which are usually manually drawn or delineated by tools with intensive user interaction as mentioned in [18]. The heavy annotation cost of training data collection hinders the advances in top-down saliency detection.

In this work, we propose a weakly supervised approach to address this issue. By weakly supervised, it means that training data with only image-level labels, each of which indicates the presence or absence of a target object in an image, are provided. Image-level labels are collected more efficiently than pixel-level ones, so the annotation cost is substantially reduced. Compared to existing weakly supervised approaches, *e.g.* [10], our approach carries out top-down saliency detection based on *convolutional neural networks* (CNNs) [27]. CNNs have demonstrated the effectiveness in joint visual feature extraction and nonlinear classifier learning. With CNNs, the highly nonlinear mapping between images and their saliency maps are better modeled. On the other hand, the sub-optimal hand-crafted features are replaced with the better features learned automatically by CNNs. Therefore, saliency maps of higher quality can be generated. Unlike most top-down saliency approaches that generate down-sampled saliency maps due to the computational issue, our approach can generate full-resolution maps, and be applied to the tasks where resolution matters.

Our approach is developed based on the observation: For a learned classifier that separates object images of a target category from the rest, it tends to have a high prediction confidence if the irrelevant background region of an object image is removed. Figure 1 gives examples of this observation. The better the background areas are masked, the higher the prediction scores are. We leverage this observation to compensate for the lack of pixel-wise annotated training data in weakly supervised saliency detection. Specifically, our approach is composed of two CNN-based modules, an *image-level classifier* and a *pixel-level map generator*, as shown in Figure 2. The classifier is learned by using image-level labels available in the weakly supervised setting. It identifies the presence or absence of the target object in a given image, and propagates prediction confidence to guide the training of the pixel-level map generator. The generator is derived to compile saliency maps with which the masked training images are better predicted by the classifier.

We found that the classifier's confidence along is insufficient. The generated saliency maps often highlight only the discriminative parts of objects, and contain false alarms. Hence, our approach further explores other evidences available in weakly supervised learning. First, the background prior can be learned by referring to the background images. This

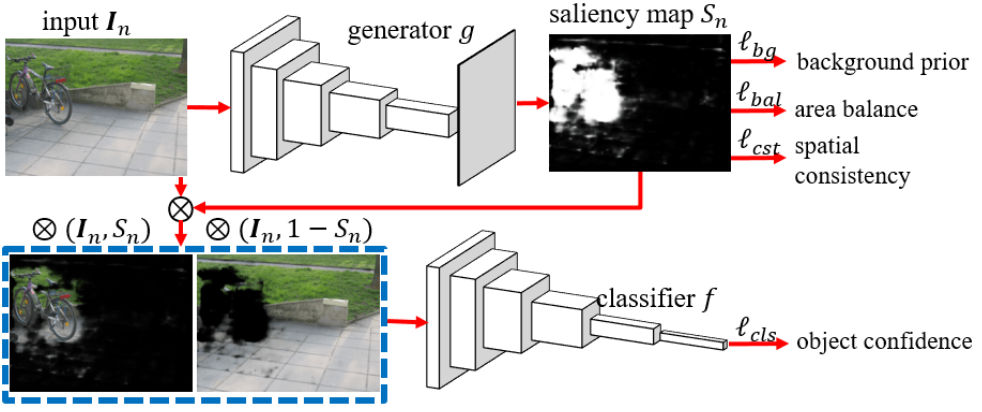


Figure 2: The overview of our proposed approach.

prior knowledge is helpful in filtering out false positives. Second, the spatial consistency in saliency maps is enforced by using graph-based regularization. It turns out that the resultant saliency maps are smooth, edge- and even object-aware. Third, the training data are often unbalanced in the sense that the background area is far larger than the foreground area. It sometimes leads to saliency maps where most pixels are predicted as non-salient especially when spatial consistency is enforced. Thus, we employ an entropy-based regularizer to avoid this unfavorable circumstance.

The main contribution of this work is to develop a general CNN-based framework for weakly supervised top-down saliency detection. It utilizes the category-driven information from the classifier to derive the generator of saliency maps. On the other hand, three additional evidences are adopted to facilitate generator training. The resulting objective function is differentiable, so the proposed approach is end-to-end trainable. Our architecture, the coupled classifier and map generator, is simple yet flexible, and can be extended to address other weakly supervised tasks such as object localization or semantic segmentation where map-like outputs can be derived by the given class label in a top-down manner. Our approach is comprehensively evaluated on the three standard benchmark datasets, Graz-02 [33], PASCAL VOC 2007 (VOC07) and PASCAL VOC 2012 (VOC12) [12], for top-down saliency detection. The results show that our approach outperforms the state-of-the-art weakly supervised approaches and many strongly supervised ones. The source code and experimental results are available from http://cvlab.citi.sinica.edu.tw/ProjectWeb/WSSD/.

2 Related work

There are many research topics related to the saliency detection. Our approach focuses on the single object saliency detection, so the eye-fixation and co-saliency methods are not covered.

Bottom-up object saliency detection Bottom-up approaches [4, 5] find objects attracting humans. Different hypothesis or priors [4, 5] are used to distinguish salient objects from background, such as global/local contrast [8, 21], focusness [22], objectness [7, 20, 22], co-segment [40] or video motion [19]. These approaches sometimes fail because the hypothesis or priors vary from object category to object category. To overcome the issue, learning-based methods, *e.g.* [24, 30, 43], are proposed to capture the concept of objects, such as the space learning [24, 30] or a random forest regressor with contrast descriptors [43]. Although

these learning-based methods often achieve better performance, there still exist limitations in bottom-up saliency detection. First, the bottom-up methods only detect the most salient objects in an image, and easily fail in the condition that multiple objects with different categories are presented in a scene. Second, they lack high-level semantic meaning, so they are difficult to be directly integrated into other tasks requiring the top-down prior.

Top-down object saliency detection Top-down saliency methods such as [9, 10, 15, 25, 46] utilize the category-specific information to learn the object concept from a set of categorized training data. These methods are confined to the pre-defined categories, so they don't suffer from the aforementioned limitations caused by the lack of category labels. Although their effectiveness has been demonstrated, these top-down methods require pixel-wise annotated training data, and result in a high annotation cost. The pioneering work by Cholakkal *et al*. [10] tackled this issue by formulating saliency detection as a weakly supervised learning problem with image-level labels.

The proposed approach also carries out top-down saliency detection in a weakly supervised fashion. The major difference between our approach and Cholakkal *et al*.'s approach [10] is that the CNN-based architecture is leveraged in our approach. Therefore, engineered features are replaced by the features learned to optimize the objective of saliency detection. Much better performance can be achieved. In addition to less costly annotation and good performance, our approach can efficiently produce full-resolution saliency maps without the extra steps for image down-sampling and map up-sampling or superpixel computing. In the state-of-the-art weakly and strongly supervised methods such as [9, 10, 25, 46], the features are computed on superpixels or over a grid to reduce the complexity. The extra quantization procedure may induce performance degradation. The CNN-based strongly supervised method [15] requires additional exemplars for both training and testing, resulting in high training and test computational cost. In addition, the additional disk storage and memory are required for the additional dataset to search the exemplars.

CNN-based weakly supervised learning Learning CNNs in a weakly supervised manner attracts much attention, and has been explored in a few computer vision tasks, such as object localization [2, 3, 23, 29] and semantic segmentation [17, 26, 38, 45]. The top-down saliency detection is related to the two tasks, and can be integrated into them, because all of them utilize the top-down knowledge. These CNN-based, weakly supervised methods for object localization and semantic segmentation often require extra information such as bounding box proposals, objectness or attention from the output of other work. Unlike them, our approach works with merely image-level labels, and achieves superior performance on saliency detection.

Top-down neural attention The methods [37, 48, 49] generate the class-specific activation maps by analyzing the neuron responses or gradients from the backward propagation of a classifier. These methods are similar to ours, but the goals are different. Our approach aims to detect the whole salient objects and keep the object shapes, while they focus on locating only the discriminative or representative parts of objects since it suffices for their respective applications, such as object localization, word attention or image captioning. Due to the dissimilar purposes, the evaluation criteria and the approach formulations are different. For example, precision@EER and the smoothness terms are used in our approach for encouraging the cover of the whole objects and preserving the object shapes. By contrast, the pointing game is used in [37, 48] to emphasize if the maximum point lies in an object.

3 The proposed approach

Our approach is introduced in this section, including the problem definition, the proposed formulation, and the optimization process. The implementation details are also included.

3.1 Problem definition

We aim at weakly supervised saliency detection with image-level annotated training data. In the stage of training, a training set of binary labels is given, $D = D_{obj} \cup D_{bg} = \{(I_n, y_n)\}_{n=1}^{N}$, where N is the number of training images. I_n is the nth training image with its label $y_n \in \{0,1\}$ indicating the presence ($y_n = 1$) or absence ($y_n = 0$) of a target object. D_{obj} and D_{bg} are the subsets of object images and background images, respectively. With D, our goal is to learn a model that accurately detects the target objects in given images in testing.

3.2 Our formulation

As shown in Figure 2, our approach is composed of two deep modules, image-level classifier $f(\cdot)$ and pixel-level map generator $g(\cdot)$. The classifier $f(\cdot)$ is learned to best separate the two-class training set D. It predicts for each I_n, and propagates the classification score to guide the training of generator $g(\cdot)$. For each I_n, the generator $g(I_n)$ estimates its saliency map S_n, which is expected to highlight the target objects if they exist. Generator $g(\cdot)$ is learned in a way where the highlighted I_n by S_n can be predicted by $f(\cdot)$ with a higher confidence. Note that the softmax layers are employed in $f(\cdot)$ and $g(\cdot)$ in the binary prediction problems. Thus, the prediction of $f(\cdot)$ and each pixel in saliency map S_n is ranged between 0 and 1. In testing, generator $g(\cdot)$ produces the saliency map $g(I)$ for an input image I.

Suppose the map generator $g(\cdot)$ is parametrized by $\mathbf{w}$. The proposed objective for training the generator $g(\cdot)$ is composed of four loss functions, and is defined by

$$\begin{aligned}\ell(\mathbf{w}) = &\sum_{I_n \in D_{obj}} \ell_{cls}(I_n;\mathbf{w}) - \lambda_{bal}\ell_{bal}(I_n;\mathbf{w}) + \lambda_{cst}\ell_{cst}(I_n;\mathbf{w}) \\ &+ \sum_{I_n \in D_{bg}} \lambda_{bg}\ell_{bg}(I_n;\mathbf{w}) + \lambda_{cst}\ell_{cst}(I_n;\mathbf{w}),\end{aligned} \tag{1}$$

where λ_{bg}, λ_{bal}, and λ_{cst} are three positive constants. The four loss functions, *i.e.* ℓ_{cls}, ℓ_{bg}, ℓ_{bal}, and ℓ_{cst}, consider the classification scores, the prediction errors in background images, the area balance of the detected salient and non-salient regions, and the spatial consistency of the saliency maps, respectively. They are defined and justified as follows.

The classification loss ℓ_{cls} guides the training of the generator by referring to the classification scores made by classifier $f(\cdot)$. Its definition on an object image I_n is given below:

$$\ell_{cls}(I_n;\mathbf{w}) = \|f(\otimes(S_n, I_n)) - 1\|^2 + \|f(\otimes(1 - S_n, I_n)) - 0\|^2, \tag{2}$$

where $S_n = g_{\mathbf{w}}(I_n)$ is the saliency map predicted by current generator $g_{\mathbf{w}}$, and $\otimes$ is the operator of element-wise multiplication. $\otimes(S_n, I_n)$ is image I_n with its estimated salient regions highlighted. The classification loss ℓ_{cls} encourages the generator $g(\cdot)$ to highlight the discriminative regions of I_n so that a high classification score $f(\otimes(S_n, I_n))$ can be obtained. The assumption behind this loss is that most discriminative regions reside in the target objects. Likewise, we hope that the background saliency map, *i.e.* $1 - S_n$, doesn't contain any object parts, so the classification score $f(\otimes(1 - S_n, I_n))$ is minimized.

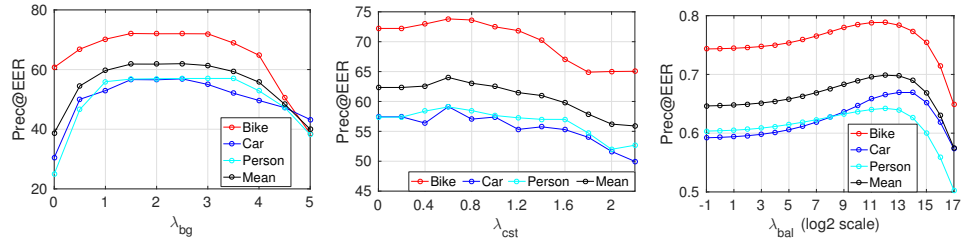


Figure 3: The performances of our approach in Prec@EER with different vales of parameter (a) λ_{bg}, (b) λ_{bal}, and (c) λ_{cst}.

The background loss ℓ_{bg} prevents the generator from detecting salient objects in a background image I_n. It is defined by

$$\ell_{bg}(I_n;\mathbf{w}) = \frac{1}{W \times H}\|S_n - Z\|^2, \tag{3}$$

where W and H are the width and the height of I_n, respectively. $Z \in \mathbb{R}^{W\times H}$ is a matrix whose elements are 0. This loss greatly reduces the false alarms in saliency detection.

The balance loss ℓ_{bal} aims to balance the areas of the predicted salient and non-salient regions in an object image I_n. In training set D, the number of the background pixels is far larger than that of the object pixels. The generator tends to produce saliency maps where the saliency values on most pixels are low. The situation even becomes worse with the use of loss ℓ_{bg}. The balance loss is then introduced as follows:

$$\ell_{bal}(I_n;\mathbf{w}) = -\bar{s}_n \log \bar{s}_n - (1-\bar{s}_n)\log(1-\bar{s}_n), \tag{4}$$

where $\bar{s}_n$ is average saliency value of map S_n. This loss function is in form of entropy, and can avoid classifying most pixels as either background or object.

The loss function ℓ_{cst} enforces the spatial consistency of the detected saliency maps by minimizing

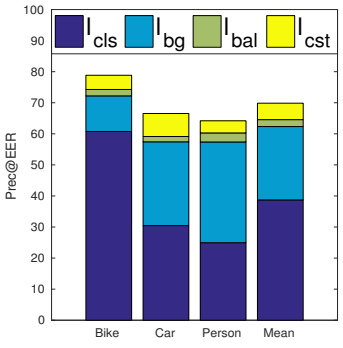
$$\ell_{cst}(I_n;\mathbf{w}) = \sum_{(i,j)\in\varepsilon} e_{i,j}\|S_n(i) - S_n(j)\|^2 = vec^\top(S_n)\, L\, vec(S_n), \tag{5}$$

where ε is the set of the edges connecting adjacent pixels, and $S_n(i)$ is the saliency value at pixel i of map S_n. $vec(\cdot)$ is the vectorization operator. The edge weight $e_{i,j}$ of two adjacent pixels i and j is defined by $\exp\big(-\frac{max(GbP_i, GbP_j)}{\sigma^2}\big)$, where GbP_i is the generalized boundary probability [28] at pixel i, and σ is set to the mean of all edge weights [36]. L is the graph Laplacian of affinity matrix $[e_{i,j}]$. This graph-based regularization term preserves discontinuity and makes the resultant saliency maps smoother.

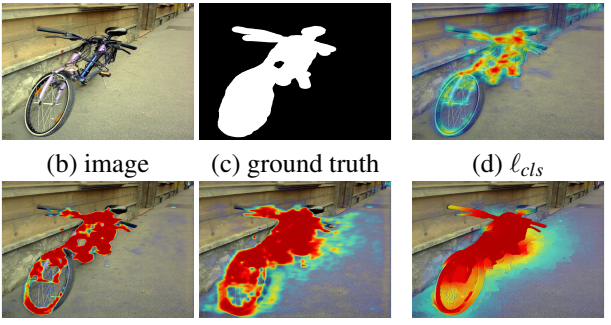
3.3 Optimization process

Directly optimizing the objective in Eq. (1) needs a large storage space due to graph Laplacian L in Eq. (5). We present a two-stage optimization process to solve this problem:

Offline optimization: During the training stage, we optimize the objective in Eq. (1) but excluding the loss function ℓ_{cst} for the training data. All the loss functions in the objective are differential and convex, so stochastic gradient descent algorithms can efficiently and



(a) Performance gain



(b) image (c) ground truth (d) ℓ_{cls}

(e) ℓ_{cls}+ℓ_{bg} (f) ℓ_{cls}+ℓ_{bg}+ℓ_{bal} (g) ℓ_{cls}+ℓ_{bg}+ℓ_{bal}+ℓ_{cst}

Figure 4: (a) The performance gains in Prec@EER obtained by adding the four loss functions, *i.e.* ℓ_{cls}, ℓ_{bg}, ℓ_{bal}, and ℓ_{cst}, one by one. (b) $\sim$ (g) Visualizations for an example.

effectively solve the optimization problem. The gradient of each loss function with respect to the optimization variables can be derived straightforward, and is omitted here.

Online optimization: During the testing stage, we solve the following objective to get a smooth saliency map for the testing image I:

$$\underset{S^*}{\operatorname{argmin}} E(S^*;S) = \|S^* - S\|^2 + \lambda_{cst} \sum_{(i,j)\in\varepsilon} e_{i,j} \|S^*(i) - S^*(j)\|^2, \tag{6}$$

where $S = g_{\mathbf{w}}(I)$ is the saliency map generated by the learned map generator $g_{\mathbf{w}}(\cdot)$. The final saliency map S^* is yielded by further considering the spatial consistency. S^* can be efficiently computed, since there exists a closed-form solution to Eq. (6), *i.e.* $vec(S^*) = (\mathbb{1} + \lambda_{cst} L)^{-1} vec(S)$, where $\mathbb{1}$ is an identity matrix.

We implemented the proposed network based on MatConvNet [42]. ResNet-50 [14] is adopted as the classifier $f(\cdot)$. It is pre-trained on ImageNet [11] and fine-tuned by using the training set D. The batch size, weight decay and momentum are set to 32, 0.0005, and 0.9, respectively. The learning rate is initially set to 0.001, and decreased by a factor of 10 every 20 epochs. In total, the learning rate is decreased 4 times, and the learning process stops after 100 epochs. The architecture of the generator network is developed based on the VGG16 [39] setting of FCN [31] with the same batch size, weight decay, and momentum. The learning rate is set to 0.00001, and fixed during training. The number of epoches is set to 100. Besides, because the classifier requires the inputs with the same size, each training image is resized to resolution 384×384 in advance.

4 Experimental results

Our approach is evaluated in this section. Firstly, the sensitivity analysis of the model parameters is conduced, and the effect of each loss function is assessed. Then, our approach is compared with the stare-of-the-art approaches.

The experiments are conducted on three datasets, Graz-02 [33], VOC07 and VOC12 [12]. Graz-02 contains three object categories (bike, car and person) and a background category, and each category has 300 images. Following the setting in [9, 10, 46], the odd numbered 150 images from each category serve as the training data, while the rest are treated as the test data. VOC07 and VOC12 are more challenging than Graz02 because of more variations and

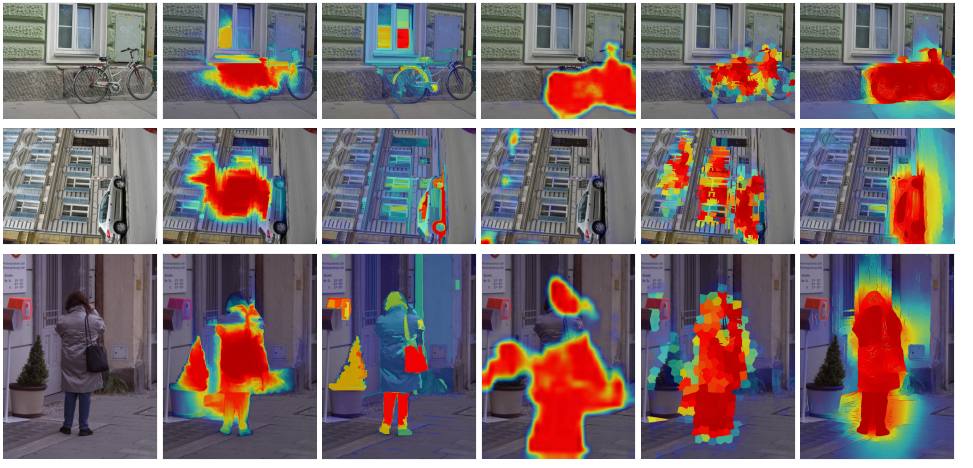


Figure 5: The saliency maps detected by our approach and the completing approaches. In the three examples, the target categories are *bike*, *car* and *person* from top to bottom.

occlusions. There are 20 categories in VOC12. Following the VOC12 setting in [15, 34], our models are trained on the training set from the classification task, and evaluated on the validation set of the segmentation task. VOC07 is the subset of VOC12, and following the setting in [9, 10, 46], our models are evaluated on the test set from the segmentation task. Data augmentation by rotation and flipping is used on these datasets, and category-specific models are trained for each category. The criterion, precision at equal error rate (Prec@EER) [9, 10, 25, 46], is adopted to measure the quality of the detected saliency maps on Graz02 and VOC07. For VOC12, the binarized Prec@EER [15] is adopted.

4.1 Model analysis

The proposed objective in Eq. (1) consists of four loss functions. Except classification loss ℓ_{cls}, the other three loss functions, ℓ_{bg}, ℓ_{bal}, and ℓ_{cst}, are associated with leading parameters, *i.e.* λ_{bg}, λ_{bal}, and λ_{cst}, respectively. We perform sensitivity analysis of the three parameters, and assess the effect of adopting these loss functions. First of all, the classification loss ℓ_{cls} is employed with its leading parameter set to 1. We add the background loss ℓ_{bg} for removing false positives in saliency maps. The performance of our approach by varying the value of the corresponding parameter λ_{bg} is shown in Figure 3a. It can be observed that ℓ_{bg} is crucial, since the performance gain by changing λ_{bg} from zero to a positive value is significant. We empirically set λ_{bg} to 2.5. Then, the third loss ℓ_{bal} is included to balance the areas of the detected salient and non-salient regions. The performance of our approach with different values of parameter λ_{bal} is similarly reported in Figure 3b. Loss ℓ_{bal} moderately enhances saliency detection. The parameter ℓ_{bal} is fixed to 0.6. Finally, the fourth loss ℓ_{cst} is introduced to make saliency maps smoother. As shown in Figure 3c, this loss remarkably helps saliency detection. Parameter λ_{cst} is set to 2^{12}. The optimal values of these parameters are roughly shared among the three object categories. We fix the parameters, *i.e.* $(\lambda_{bg}, \lambda_{bal}, \lambda_{cst}) = (2.5, 0.6, 2^{12})$, in the following experiments.

To quantify the effect of each of the four loss functions, we report the performance gains obtained by sequentially adding these losses, ℓ_{cls}, ℓ_{bg}, ℓ_{bal}, and ℓ_{cst}. The results in Figure 4a indicate that each loss function makes contribution to saliency detection for all the three

Table 1: Performance in Prec@EER (%) of different approaches on three benchmarks.

Graz02			
Group	Method	Setting	Mean
Bottom-up	MB [47]	US	48.6
	MST [41]	US	46.7
	HDCT [24]	FS	50.9
	DRFI [43]	FS	53.5
Top-down	Aldavert et al. [1]	FS	65.1
	Fukerson et al. [13]	FS	70.2
	Shape mask [32]	FS	53.2
	Yang and Yang [46]	FS	61.3
	Kocak et al. [25]	FS	70.2
	LCCSC [9]	FS	70.5
	R-ScSPM [10]	FS	72.1
	R-ScSPM [10]	WS	60.5
	Ours	WS	69.9

VOC07		
Method	Setting	Mean
Yang and Yang [46]	FS	16.7
LCCSC [9]	FS	23.4
R-ScSPM [10]	WS	18.6
Ours	WS	23.5

VOC12		
Method	Setting	Mean
Yang and Yang [46]	FS	15.6
Kocak et al. [25]	FS	40.4
He et al. [15]	FS	56.2
Oquab et al. [34]	WS	48.1
Ours	WS	50.0

object categories. To better understand the gains, an example of the detected saliency maps generated through the procedure of sequentially adding the four loss functions is given in Figure 4d $\sim$ 4g. With only the classification loss ℓ_{cls}, the target object, the bicycle here, is detected, but many false alarms occur. From Figure 4d to 4e, the background loss ℓ_{bg} is added, and helps remove most false alarms. From Figure 4e to 4f, the added balance loss ℓ_{bal} makes the saliency map sharper, especially in the region of the object. From Figure 4f to 4g, the consistency loss ℓ_{cst} makes the saliency map much smoother and better preserves the object boundary.

4.2 Comparison with the state-of-the-arts

We compare our proposed method with the state-of-the-art methods and report the results in Table 1, where *setting* denotes the supervision condition of training data including US (unsupervised), WS (weakly supervised), and FS (fully supervised). On Graz-02, the bottom-up methods [24, 41, 43, 47] identify salient objects without using any prior of a target category. Despite the broad applicability, they do not perform very well for category-specific saliency detection. Instead, top-down methods [1, 9, 13, 25, 32, 46] learn the discriminative information by using pixel-wise annotated training data, and get much better performance. However, collecting such training data is costly. R-ScSPM [10] and our method adopt the weakly supervised setting, and can work with image-wise annotated training sets. Our method leverages multiple evidences and integrates them into a CNN-based network architecture. It turns out that our method outperforms R-ScSPM [10] by a large margin around 9.4% in Prec@EER. It is worth mentioning that our method achieves a comparable performance to the state-of-the-art fully supervised methods, and even beats some of them. In more challenging datasets, VOC07 and VOC12, our method also performs well. In VOC07, our method can outperform all the baselines, including the fully supervised [9, 46] and the weakly supervised method [10], because CNN can learn more powerful features to identify the target. In VOC12, our method can also outperform the non-CNN fully supervised [25, 46] and CNN weakly supervised [34] methods. In [34], the authors aim to localize objects with the trained classifier. They only search the discriminative object parts, so their method can't identify the whole salient target. Although our performance is lower than that

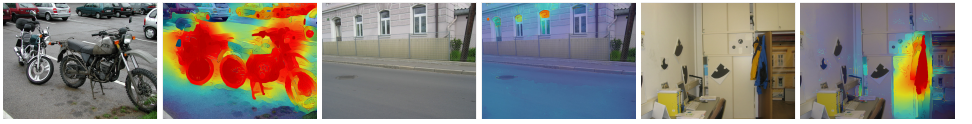


Figure 6: Some failure cases show the positive alarms. The first, third and fifth images are the background images from categories *bike*, *car*, and *person*, respectively, and their corresponding saliency maps are shown in the second, fourth and sixth images, respectively.

of [15], the framework in [15] requires the pixelwise annotations to learn the parameters. However, we don't need any pixelwise object masks, and the annotation cost can be reduced.

To gain insight into the quantitative results, Figure 5 shows some detected saliency maps by different approaches. The bottom-up approach, MB [47] and DRFI [43], have many false positives due to the lack of category-specific information in training data. False positives are prone to happen in the regions of high contrast, such as windows, bags, and clothes. Compared to MB and DRFI, the top-down methods [25, 46] can detect satisfactory saliency maps. However, they still have a few limitations. First, the adopted engineered features are less discriminative. Thus, there are still a few false positive. Second, their features are extracted from a patch [46] or a superpixel [25] to reduce the complexity. The finer structures are not well preserved, and may have the block effect. Our proposed method is developed upon CNNs and can be efficiently computed in testing. It does not suffer from the aforementioned issues and can produce saliency maps of higher quality.

We show some failure cases by our approach in Figure 6. Most failure cases made by our approach are caused by the high similarity between target objects and the background (including objects of non-target categories). The motorbikes in the first image have the appearance similar to bikes, so they are detected as salient. In the third image, the windows of the buildings look like those of cars. Our approach does not explore contextual information and leads to false detection. In the last case, clothes and jackets are usually present with persons. When they are present alone, false alarms occur.

5 Conclusions

We have presented a novel approach that carries out top-down saliency detection in a weakly supervised manner. Our approach is composed of two deep modules, an image-level classifier and a pixel-level saliency map generator. During training, the knowledge of the class labels is propagated from the classifier to guide the training of the generator. The training process is further regularized by leveraging other evidences available in weakly supervised learning, including the background prior, area balance between the salient and non-salient regions, and spatial consensus, with which the effect of overfitting can be alleviated. We comprehensively analyze the effect of introducing each adopted loss function, and show that these loss functions are useful and are not sensitive to the parameters. The experimental results demonstrate that our method outperforms the existing weakly supervised methods and is comparable to the state-of-the-art fully supervised methods. In future, we plan to generalize this approach to deal with multi-label cases so that it can be applied to other target-oriented tasks such as object localization or semantic segmentation.

Acknowledgments. This work was supported in part by grants MOST 105-2221-E-001-030-MY2, MOST 105-2218-E-001-006, MOST 105-2218-E-002-011 and MOST 105-2218-E-002-032.

References

[1] D. Aldavert, A. Ramisa, R. L. de Mantaras, and R. Toledo. Fast and robust object segmentation with the integral linear classiifier. In *CVPR*, 2010.

[2] A. Bearman, O. Russakovsky, V. Ferrari, and F.-F. Li. What's the point: Semantic segmentation with point supervision. In *ECCV*, 2016.

[3] H. Bilen and A. Vedaldi. Weakly supervised deep detection networks. In *CVPR*, 2016.

[4] A. Borji, M.-M. Cheng, H. Jiang, and J. Li. Salient object detection: A survey. *arXiv*, 2014.

[5] A. Borji, M.-M. Cheng, H. Jiang, and J. Li. Salient object detection: A benchmark. *TIP*, 2015.

[6] K.-Y. Chang, T.-L. Liu, H.-T. Chen, and S.-H. Lai. From co-saliency to cosegmentation: An efficient and fully unsupervised energy minimization model. In *CVPR*, 2011.

[7] K.-Y. Chang, T.-L. Liu, H.-T. Chen, and S.-H. Lai. Fusing generic objectness and visual saliency for salient object detection. In *ICCV*, 2011.

[8] M.-M. Cheng, N. J. Mitra, X. Huang, P. H. S. Torr, and S.-M. Hu. Global contrast based salient region detection. *TPAMI*, 2014.

[9] H. Cholakkal, D. Rajan, and J. Johnson. Top-down saliency with locality-constrained contextual sparse coding. In *BMVC*, 2015.

[10] H. Cholakkal, J. Johnson, and D. Rajan. Backtracking ScSPM image classifier for weakly supervised top-down saliency. In *CVPR*, 2016.

[11] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A preview of a large-scale hierarchical database. In *CVPR*, 2009.

[12] M. Everingham, L. Van Gool, C. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (VOC) challenge. *IJCV*, 2010.

[13] B. Fulkerson, A. Vedaldi, , and S. Soatto. Class segmentation and object localization with superpixel neighborhoods. In *ICCV*, 2009.

[14] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[15] S. He, R. Lau, and Q. Yang. Exemplar-driven top-down saliency detection via deep association. In *CVPR*, 2016.

[16] S. Hong, T. You, S. Kwak, and B. Han. Online tracking by learning discriminative saliency map with convolutional neural network. In *ICML*, 2015.

[17] Q. Hou, P. Dokania, D. Massiceti, Y. Wei, and M.-M. Cheng P. H. S. Torr. Mining pixels: Weakly supervised semantic segmentation using image labels. *arXiv*, 2016.

[18] K.-J. Hsu, Y.-Y. Lin, and Y.-Y. Chuang. Augmented multiple instance regression for inferring object contours in bounding boxes. *TIP*, 2014.

[19] C.-R. Huang, Y.-J. Chang, Z.-X. Yang, and Y.-Y. Lin. Video saliency map detection by dominant camera motion removal. *TCSVT*, 2014.

[20] Y. Jia and M. Han. Category-independent object-level saliency detection. In *ICCV*, 2013.

[21] H. Jiang, J. Wang, Z. Yuan, T. Liu, and N. Zheng. Automatic salient object segmentation based on context and shape prior. In *BMVC*, 2011.

[22] P. Jiang, H. Ling, J. Yu, and J. Peng. Salient region detection by UFO: Uniqueness, focusness and objectness. In *ICCV*, 2013.

[23] V. Kantorov, M. Oquab, M. Cho, and I. Laptev. Contextlocnet: Context-aware deep network models for weakly supervised localization. In *ECCV*, 2016.

[24] J. Kim, D. Han, Y.-W. Tai, and J. Kim. Salient region detection via high-dimensional color transform. In *CVPR*, 2014.

[25] A. Kocak, K. Cizmeciler, A. Erdem, and E. Erdem. Top down saliency estimation via superpixel-based discriminative dictionaries. In *BMVC*, 2014.

[26] A. Kolesnikov and C. Lampert. Seed, expand and constrain: Three principles for weakly-supervised image segmentation. In *ECCV*, 2016.

[27] A. Krizhevsky, I. Sutskever, and G. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.

[28] M. Leordeanu, R. Sukthankar, and C. Sminchisescu. Generalized boundaries from multiple image interpretations. *TPAMI*, 2014.

[29] D. Li, J.-B. Huang, Y. Li, S. Wang, and M.-H. Yang. Weakly supervised object localization with progressive domain adaptation. In *CVPR*, 2016.

[30] S. Li, H. Lu, Z. Lin, X. Shen, and B. Price. Adaptive metric learning for saliency detection. *TIP*, 2015.

[31] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional models for semantic segmentation. In *CVPR*, 2015.

[32] M. Marszalek and C. Schmid. Accurate object recognition with shape masks. *IJCV*, 2012.

[33] A. Opelt, A. Pinz, M. Fussenegger, and P. Auer. Generic object recognition with boosting. *TPAMI*, 2006.

[34] M. Oquab, L. Bottou, I. Laptev, and J. Sivic. Is object localization for free? - weakly-supervised learning with convolutional neural networks. In *CVPR*, 2015.

[35] Z. Ren, S. Gao, L.-T. Chia, and I. W.-H. Tsang. Region-based saliency detection and its application in object recognition. *TCSVT*, 2014.

[36] C. Rother, V. Kolmogorov, and A. Blake. "GrabCut" - Interactive foreground extraction using iterated graph cuts. *TOG*, 2004.

[37] R. Selvaraju, A. Das, R. Vedantam abd M. Cogswell, D. Parikh, and Dhruv Batra. Grad-cam: Why did you say that? visual explanations from deep networks via gradient-based localization. In *NIPS Worhshop*, 2016.

[38] W. Shimoda and K. Yanai. Distinct class-specific saliency maps for weakly supervised semantic segmentation. In *ECCV*, 2016.

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

[40] C.-C. Tsai, X. Qian., and Y.-Y. Lin. Segmentation guided local proposal fusion for co-saliency detection. In *ICME*, 2017.

[41] W.-C. Tu, S. He, Q. Yang, and S.-Y. Chien. Real-time salient object detection with a minimum spanning tree. In *CVPR*, 2016.

[42] A. Vedaldi and K. Lenc. MatConvNet – convolutional neural networks for matlab. In *ACMMM*, 2015.

[43] J. Wang, H. Jiang, Z. Yuan, M.-M. Cheng, X. Hu, and N. Zheng. Salient object detection: A discriminative regional feature integration approach. *IJCV*, 2016.

[44] W. Wang, J. Shen, and F. Porikli. Saliency-aware geodesic video object segmentation. In *CVPR*, 2015.

[45] Y. Wei, X. Liang, Y. Chen, X. Shen, M.-M. Cheng, J. Feng, Y Zhao, and S. Yan. Stc: A simple to complex framework for weakly-supervised semantic segmentation. *TPAMI*, 2016.

[46] J. Yang and M.-H. Yang. Top-down visual saliency via joint CRF and dictionary learning. In *CVPR*, 2012.

[47] J. Zhang, S. Sclaroff, Z. Lin, X. Shen, B. Price, and R. Měch. Minimum barrier salient object detection at 80 fps. In *ICCV*, 2015.

[48] J. Zhang, Zhe Lin, J. Brandt, X. Shen, and S. Sclarff. Top-down neural attention by excitation backprop. In *ECCV*, 2016.

[49] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Learning deep features for discriminative localization. In *CVPR*, 2016.

[50] L. Zhu, Z. Chen, X. Chen, and N. Liao. Saliency & structure preserving multi-operator image retargeting. In *ICASSP*, 2016.