Weakly Supervised Salient Object Detection by Learning A Classifier-Driven Map Generator

Kuang-Jui Hsu, Yen-Yu Lin, *Member, IEEE,* and Yung-Yu Chuang, *Member, IEEE*

***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 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 collaborative CNN 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 by which the 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 exploring other cues, including the background prior, superpixel- and object proposal-based evidence. The background prior is introduced to reduce false positives. Evidence from superpixels helps preserve sharp object boundaries. The clue from object proposals improves the integrity of highlighted objects. These different types of cues greatly regularize the training process and reduces the risk of overfitting, which happens frequently when learning CNN models with few training data. Experiments show that our method achieves superior results, even outperforming fully supervised methods.**

***Index Terms*—Top-down object saliency detection, convolutional neural networks, weakly supervised learning.**

I. Introduction

OBJECT saliency detection has been an active topic in the fields of image processing and computer vision for decades. The detected saliency maps highlight the regions of objects attracting people. They are crucial to various applications such as image retargeting [1], visual tracking [2], object segmentation [3], [4] and object recognition [5], because they can indicate objects of interest and mask out irrelevant background.

Following the previous studies of top-down saliency detection [6]–[11] object saliency detection methods can be roughly divided into the bottom-up and the top-down groups. Bottom-up methods rely on merely the information derived from images alone for saliency 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 the background is subtle. By contrast, top-down approaches, e.g., [6]–[11], are category-aware. They utilize the prior knowledge about the target object category, such as object segment or bounding box annotations, for saliency detection, and suffer less from the aforementioned limitation. However, the top-down methods need training data in the form of pixel-wise annotations, indicating whether a pixel belongs to the target object category, which are usually manually drawn or delineated by tools with intensive user interaction as mentioned in previous work [12], [13]. The heavy annotation cost of training data collection hinders the advances in top-down saliency detection.

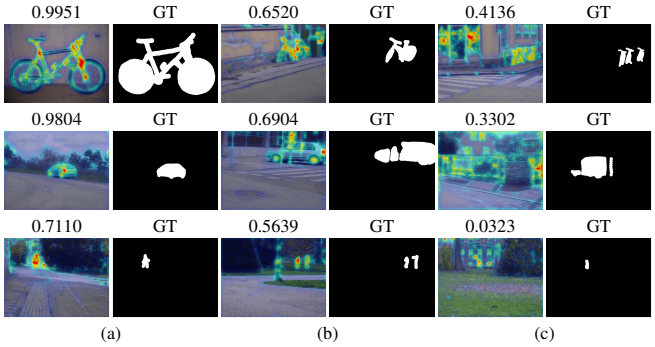

Fig. 1. Examples of the detected saliency maps and their ground truth annotations for the categories *bike* (top row), *car* (middle row), and *person* (bottom row). On the top of each map, we show the score by applying the classifier to the image with its non-salient regions removed. For images in (a), the saliency maps are of high quality and their classification scores are also very high. For images in (b), the map quality is worse and the scores are lower. Finally, for images in (c), the low-quality saliency maps lead to even lower classification scores since more irrelevant background is retained and it could disturb the classifier. It is clear that the better the non-salient areas are removed, the higher the classification scores are.

In this paper, we propose a weakly supervised approach for addressing this issue. Our weakly supervised method only requires training data with image-level labels, each of which indicates the presence or absence of a target object in an image. Image-level labels can be collected more efficiently than pixel-level ones, so the annotation cost is substantially reduced. Even better, many such annotations have already been collected for other problems such as image classification. Compared to the existing weakly supervised approaches, e.g., [9], [11], our approach carries out top-down saliency detection based on *convolutional neural networks* (CNNs) [14]. 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. At the same time, the sub-optimal hand-

K.-J. Hsu is with the Research Center for Information Technology Innovation, Academia Sinica, Taipei 115, Taiwan and the Department of Computer Science and Information Engineering, Nation Taiwan University, Taipei 106, Taiwan. E-mail: kjhsu@citi.sinica.edu.tw

Y.-Y. Lin is with the Research Center for Information Technology Innovation, Academia Sinica, Taipei 115, Taiwan. E-mail: yylin@citi.sinica.edu.tw

Y.-Y. Chuang is with the Department of Computer Science and Information Engineering, Nation Taiwan University, Taipei 106, Taiwan and the Research Center for Information Technology Innovation, Academia Sinica, Taipei 115, Taiwan. E-mail: cyy@csie.ntu.edu.tw

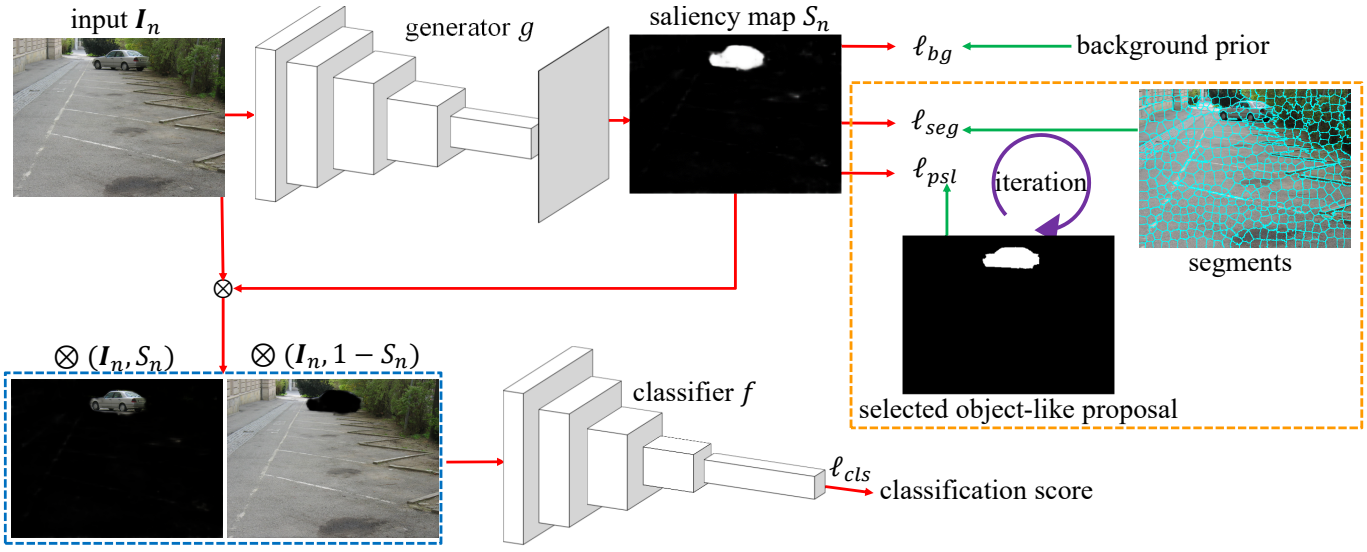

Fig. 2. The overview of our approach. The classifier f distinguishes images of a target class from the rest. It propagates the classification information via the loss function ℓ_{cls} to train the generator g, which compiles saliency maps so that the masked images can be predicted by f with higher confidence. The other three loss functions, ℓ_{bg}, ℓ_{seg} and ℓ_{psl}, explore cues from the background prior, superpixels, and object proposals, respectively. They are introduced for generating high-quality saliency maps.

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 is suitable for the tasks where resolution matters.

Our approach is developed based on the following observation. For a classifier that separates object images of a target category from the rest, it tends to have a high prediction confidence if the irrelevant background of an object image is removed. Fig. 1 gives some examples of this observation. The better the background areas are masked out, 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 Fig. 2. The classifier is learned by using image-level labels available in weakly supervised learning. It identifies the presence or absence of the target object in an 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.

A similar observation is also leveraged by Cholakkal et al. [11] but with two major differences. First, our method utilizes this observation in the *training* stage to learn the generator while Cholakkal et al. [11] use it in the *inference* stage as post-processing by fusing the best bottom-up saliency maps generated by other methods. This difference makes our method more efficient than their method [11] in inference since our method only needs to apply the generator trained based on this observation to the test images while theirs has to invoke several other map generators and perform fusion in the inference stage. Second, our method leverages this observation to derive the generator, which directly *learns* the saliency maps from *raw input images*. The method in [11] uses this observation to *fuse* the *saliency maps* generated by other existing methods in the domain of *the extracted features* rather than images or maps themselves. Thus, the performance of the method in [11] highly depends on the quality of the maps generated by other methods and may suffer from the information loss caused by the conversion from an image/map to the features.

The collaboration between the image classifier and the map generator enables weakly supervised top-down saliency detection. However, the collaboration alone is insufficient to result in saliency maps of high quality. The generated saliency maps often have false alarms, are blurred especially near object boundaries, and highlight only discriminative object parts. Hence, our approach further explores other evidence to address these issues. First, the background prior is learned by referring to the background images. This prior knowledge is helpful in filtering out false positives. Second, we compute *superpixels*, which reveal two important clues for saliency detection: 1) Most object boundaries are discovered; 2) the pixels within a superpixel tend to belong to the object or the background all together. We leverage the clues to make saliency maps sharper while removing noise. Third, we generate object-like *proposals*. The evidence jointly explored by saliency detection and proposal selection helps recover non-discriminative object parts, making the whole objects completely highlighted in the saliency maps.

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. In addition, three additional types of evidence are adopted to enhance generator training. The resulting objective function is differentiable, so the proposed approach is end-

to-end trainable. Our architecture, the coupled CNN-based classifier and map generator, is simple yet flexible. It can be extended to address other weakly supervised tasks such as object localization or semantic segmentation where map-like outputs are derived from the given class labels in a top-down manner. Our approach is comprehensively evaluated on three standard benchmark datasets for top-down saliency detection, including Graz-02 [15] and PASCAL VOC-07/12 [16]. The results show that our approach outperforms the state-of-the-art weakly supervised approaches and many fully supervised ones in both accuracy and efficiency.

II. Related work

Saliency detection is an active topic in image processing and computer vision, and has several important branches, such as singe-image object saliency detection, object co-saliency detection, and eye-fixation. Our review mainly focuses on singe-image object saliency detection because it is the most relevant to our proposed method.

A. Bottom-up object saliency detection

Bottom-up object saliency detection [17] receives much research attention owing to superior computational efficiency and less requirement of training data. As discussed in the survey paper [17], bottom-up approaches find objects attracting humans by referring to different category-independent object observations or priors to distinguish salient objects from the background, such as center-surround contrast [18], [19], global/local contrast [20], [21], focusness [22], objectness [22]–[24]. These approaches sometimes fail because the observations or priors vary from object category to object category. To overcome the issue, learning-based methods, e.g., [25]–[27], were proposed to capture the concept of objects, such as the space learning [25], [27] or a random forest regressor with contrast descriptors [26]. Recently, more and more researches [28]–[36] utilize CNNs to carry out the tasks of bottom-up object saliency detection in different ways, such as multi-level feature aggregation [33], uncertain convolutional feature learning [34], global context and local context integration [35], and contour-saliency conversion [36].

Wang et al. [32] proposed a two-stage method to learn a bottom-up saliency model by using image-level labeled training data. At the first stage, the foreground inference network with the proposed global smooth pooling is trained on the ImageNet dataset. At the second stage, a self-training scheme is applied by taking as input the pseudo ground truth, which is initialized at the first stage and iteratively refined by using CRF. On the contrary, our method is designed for top-down saliency detection. In addition, our method is non-iterative and end-to-end trainable, thereby leading to higher training efficiency.

Despite the effectiveness, learning-based approaches to bottom-up saliency detection have limited performance. First, the definition of salient objects is ambiguous especially when multiple objects are presented in an image. Bottom-up methods only detect the most salient object in an image, and probably fail in the condition that multiple objects of different categories are presented in a scene. Second, they lack high-level semantic meaning, so it is difficult to integrate them into the optimization process of other tasks requiring the top-down prior. Thus, they are usually used for pre-processing.

B. Top-down object saliency detection

Top-down saliency methods such as [6]–[9], [11] utilize the category-specific information to learn the object concept from a set of categorized training data. These methods are confined to pre-defined categories, so they don't suffer from the aforementioned limitations caused by the lack of category labels. Yang and Yang [6] proposed a method for top-down saliency detection by jointly learning conditional random fields and a dictionary. Kocak et al. [7] computed the first and second order statistics and objectness on superpixels to distinguish target objects from the background. Cholakkal et al. [8] proposed the *locality-constrained contextual sparse coding* (LCCSC) method for top-down saliency detection. He et al. [10] proposed an exemplar-based method with the strongly supervised CNNs guided by the selected exemplars for both training and testing. Despite the effectiveness, these top-down methods require pixel-wise annotated training data, and result in a high annotation cost. The pioneering work by Cholakkal et al. [9], [11] tackled this issue by formulating saliency detection as a weakly supervised learning problem where only image-level labels are provided.

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 [9] 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 as shown in the experiments.

Cholakkal et al. [11] later extended their work by using CNN features and employing two-step post-processing, bottom-up saliency map fusion and multi-scale superpixel-averaging. Their method achieves very satisfactory performance. Compared with their method, our method has the following two advantages. First, the method in [11] is derived based on the spatial pyramid pooling (SPP) and the formulation of the linear SVM. Thus, feature extraction and saliency detection are treated as separate steps. In contrast, our method jointly learns the CNN features and estimates saliency maps through end-to-end optimization. Second, the method in [11] relies on superpixels and multiple saliency maps produced by other off-the-shelf methods at the inference stage. Therefore, its performance depends on the saliency maps yielded by other methods and its efficiency is worse. In contrast, our method carries out saliency detection by simply applying the learned CNN model to test images. It requires neither superpixel extraction nor saliency map fusion, thereby leading to much higher efficiency. In addition, our method outperforms the method in [11] if the two-step post-processing is turned off.

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

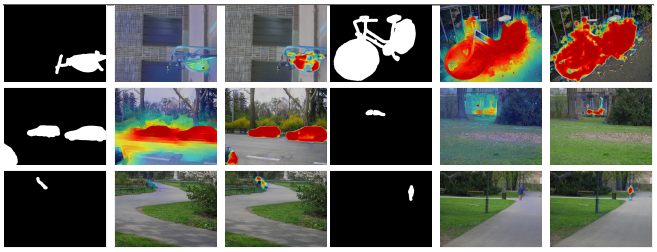

Fig. 3. Comparison between the proposed method and our prior method. Each example consists of the ground truth (GT) and two saliency maps, generated by our prior work [37] and this work, respectively. Examples from three object categories, including *bike*, *car*, and *person*, are displayed in the three rows, respectively.

map up-sampling or superpixel computing. In the non-CNN-based state-of-the-art methods for either weakly or strongly supervised saliency detection such as [6]–[9], [11], the features are computed on superpixels or over a grid to reduce the complexity. The extra quantization procedure may induce performance degradation. In the CNN-based method [10], a sliding window scheme is used to produce the saliency map of an image. Thus, multiple forward passes are required and they lead to a high computational cost. Instead, in our method, one forward pass is sufficient to perform saliency detection. Our method is 142 times faster than the CNN-based method [10] as reported in the experiments.

This work shares the same network architecture with our prior work [37], namely an architecture consisting of two collaborative CNN modules, the map generator and the image classifier, for weakly supervised saliency detection. The image classifier propagates the image-level information to train the map generator. The paper provides significant improvements by enhancing the loss function and the optimization procedure to address the limitations of the prior work. First, the map generator is learned by referring to prediction scores made by the classifier. Therefore, our prior method tends to detect only the discriminative parts of salient objects. The less discriminative regions of salient objects are sometimes missing. In addition, the prior method is prone to miss small salient objects and the detected saliency maps are blurred, especially near object boundaries. We address these limitations by integrating segmentation- and object proposal-guided evidence into the loss function. Thus, this work can better recover the whole salient regions, discover small objects and preserve object boundaries. We show some detected saliency maps by the prior work [37] and this work in Fig. 3 for comparison. It is clear that the above-mentioned limitations are properly addressed by this work. Second, a two-stage optimization procedure is adopted for saliency detection in our prior work. The second stage is used to enforce the smoothness of saliency maps. Although improving quality, the stage represents the computational bottleneck in the framework. In this paper, we add the information extracted from segmentation and object proposals to the loss function for model training. It significantly improves the quality. Thus, post-processing is no longer required. It turns out that the proposed method is about 350 times faster than the prior method [37].

C. 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 [38]–[43] and semantic segmentation [44]–[47]. Top-down saliency detection is related to the two tasks, and can be integrated into them, because all of them utilize the top-down, class-specific knowledge.

Among them, the object localization methods in [42], [43] are the most similar to our approach, because they generate saliency maps, too. The approach in [43] produces the class-specific score maps, which are aggregated into a score vector by using global max-pooling to optimize the multi-class logistical loss. However, using max-pooling is prone to find merely the discriminative parts of an object rather than the whole object. Zhou et al. [42] replaced global max-pooling with global average pooling to alleviate this problem, but the global average pooling tends to over-estimate object regions because it takes all the activations into account. Both methods in [43] and [42] only produce coarse saliency maps to save computation. The spatial structure and the object boundaries are also missing because of the use of the pooling operators. In our work, the generated maps are full-resolution, so the spatial structure can be maintained. With the aid of superpixels and object-like proposals, object boundaries and the non-discriminative object parts can be well discovered in our approach.

It is worth mentioning that semantic segmentation and top-down saliency detection are highly correlated but different. First, semantic segmentation aims to generate object segments of classes of interest. It is a task of dense or pixel-wise classification. Thus, the order of the class probabilities on each pixel is crucial, and the segmentation results are discrete. In contrast, top-down saliency detection produces the probability map encoding the occurrence likelihood of salient objects. The values in the resultant saliency maps are real-valued. Second, according to the task goals, semantic segmentation is often measured by IoU (intersection over union) and pixel-wise accuracy rates, while top-down saliency detection is usually evaluated by jointly considering precision and recall. Third, according to the evaluation metrics, CNN-based methods for semantic segmentation often employ loss functions based on softmax or other classification-based criteria. In contrast, methods for top-down saliency detection, including ours, often use the L_2 or L_1 norm loss, and take the absolute magnitudes of the saliency maps into account.

D. Top-down neural attention

Different from top-down object saliency detection, the methods in [42], [48]–[52] analyze the neuron responses or gradient of a classifier to generate class-specific activation maps. In [48], [49], the partial derivatives of neuron activations from error backward propagation are computed to highlight important image regions. In [50], a feedback loop is proposed to infer the activation of hidden layer neurons, and the feedback mechanism outputs the top-down attention which can identify discriminative object parts. Zhou et al. [42] proposed *class activation mapping* (CAM), which substitutes an average

pooling layer for the fully-connected layer. Their method helps generate coarse maps highlighting objects. Based on the winner-take-all principle and the probabilistic formulation, Zhang et al. [51] focused on generating highly discriminative attention maps. These methods aim to identify discriminative regions for a given class, and most of them are applied to the classification or localization tasks where detecting precise object boundaries and the whole objects are not necessary.

Methods discussed above have several limitations. First, these methods depend on the neuron responses of a classifier, and the activation maps are usually smaller than the input images. Therefore, test images must be resized to meet the learned models, and outputs are then resized back to original sizes. The step of image resizing often results in object distortion and makes it difficult for the attention maps to preserve object boundaries. Second, these methods find only discriminative object parts, and neglect non-discriminative but salient parts, so they cannot well identify complete objects. Third, they perform both the forward and backward propagation for each test image, so the computational cost is high. In our framework, the *fully convolutional networks* (FCN) [53] architecture is adopted for the generator, and image resizing is not required. Thus, distortion seldom happens. Superpixel segmentation and object proposals are extracted to regularize the training of CNNs. The evidence from superpixels and proposals helps preserve object boundaries and discover non-discriminative object parts. Furthermore, our approach is more efficient since it needs just one forward pass for detecting the saliency map of an input image.

III. The proposed approach

In this section, we first give the problem definition. Then, the proposed formulation and its optimization are described. Finally, the implementation details of our approach are provided.

A. 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 testing images.

B. Our formulation

As shown in Fig. 2, our approach is composed of two CNN modules, the image-level classifier $f(\cdot)$ and the 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 the generator $g(\cdot)$. For each I_n, the generator $g(I_n)$ estimates its saliency map S_n, which highlights the target objects if they exist. The 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 proposed method uses the sigmoid function as the activation functions in the last layers of both $f(\cdot)$ and $g(\cdot)$. Thus, the prediction of $f(\cdot)$ and each pixel in the saliency map S_n ranges between 0 and 1. In the phase of testing, the generator $g(\cdot)$ produces the saliency map $g(I)$ for an input image I with one forward pass.

The classifier $f(\cdot)$ is a deep model derived to separate the two-class training set D. Once the classifier $f(\cdot)$ is obtained, we focus on learning the map generator $g(\cdot)$. Suppose the 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

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

where λ_{bg}, λ_{seg}, and λ_{psl} are constants for weighting losses. M_n is the set of the superpixels extracted in image I_n. O_n is the selected object proposal for I_n. The four loss functions, i.e., ℓ_{cls}, ℓ_{bg}, ℓ_{seg}, and ℓ_{psl}, estimate the quality of saliency maps by considering the classification scores, the prediction errors in the background images, the superpixel-wise consistency of the saliency maps, and the difference between the saliency map and the selected object proposal, respectively. They are defined and justified as follows.

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

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

where $S_n = g_{\mathbf{w}}(I_n)$ is the saliency map predicted by the current generator $g_{\mathbf{w}}$, and $\otimes$ is the operator of element-wise multiplication. Thus, $\otimes(S_n, I_n)$ is the image I_n with its estimated salient regions highlighted. The classification loss $\ell_{cls}(I_n; \mathbf{w})$ 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 function is that most discriminative regions reside in the target objects. We also consider the symmetric counterpart. Namely, the non-salient areas, $1 - S_n$, should not contain any object parts. Thereby, the classification score $f(\otimes(1-S_n, I_n))$ is minimized.

2) Background loss ℓ_{bg}: It 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 false alarms in saliency detection.

3) Segmentation-based loss ℓ_{seg}: The classification loss ℓ_{cls} and the background loss ℓ_{bg} are designed to identify the regions that are classified with high confidence as foreground and background, respectively. Therefore, the two loss functions often seek the discriminative object parts and exclude the non-salient regions whose appearance is similar to the background images. Using the two loss functions alone is insufficient to preserve object boundaries, and some noises are present in the saliency maps.

We address these issues by utilizing clues from segmentation. For each image in D, we decompose it into superpixels, which have the following two properties helpful for saliency detection. First, pixels within the same superpixel tend to belong to either a salient object or the background all together. Second, object boundaries often coincide with boundaries between superpixels from over-segmentation. The former property can be used to filter out noises in a superpixel-wise manner, while the latter can be leveraged to preserve object boundaries and generate sharper saliency maps. Specifically, the segmentation-based loss for the image I_n is given below:

$$\begin{aligned}\ell_{seg}(I_n, M_n; \mathbf{w}) = &\frac{1}{W \times H} \sum_{p \in M_n} \sum_{i \in p} \\ &\left[\frac{\sum_{j \in p} S_n(j)}{|p|} > 0.5\right] \|S_n(i) - 1\|^2 \\ &+ \left[\frac{\sum_{j \in p} S_n(j)}{|p|} \leq 0.5\right] \|S_n(i) - 0\|^2,\end{aligned} \tag{4}$$

where M_n is the set of superpixels of I_n, $[\cdot]$ is the indicator function, $S_n(i)$ is the saliency value of I_n at pixel i, and $|p|$ is the size of the superpixel p. $\frac{\sum_{j \in p} S_n(j)}{|p|}$ is the average saliency value of the superpixel p. In Eq. (4), we maximize the saliency value of a pixel if it belongs to a superpixel whose average saliency value is larger than 0.5, otherwise we minimize it. Eq. (4) can be expressed equivalently as the following matrix form:

$$\ell_{seg}(I_n, M_n; \mathbf{w}) = \frac{1}{W \times H} \|S_n - G_n\|^2, \tag{5}$$

where $G_n \in \{0, 1\}^{W \times H}$ is a mask decided by average saliency values of superpixels, and is defined as

$$G_n(i) = \begin{cases} 1, & \text{if } \frac{\sum_{i \in p} S_n(i)}{|p|} > 0.5, \\ 0, & \text{otherwise}, \end{cases} \tag{6}$$

where p is the superpixel containing the pixel i.

4) Proposal loss ℓ_{psl}: When the three aforementioned loss functions, i.e., ℓ_{cls}, ℓ_{bg}, and ℓ_{seg}, are used for saliency detection, some salient objects cannot be detected completely because none of the three loss functions encourages the detection of the non-discriminative parts of salient objects. It leads to incomplete objects or objects with holes in the resultant saliency maps. This problem can be alleviated by taking the clue derived from objectness into account. To this end, we compile a pool of object proposals for each image $I_n \in D$ by using any existing, unsupervised algorithm for proposal generation. We pick the proposal that is the most consistent with the saliency map, and further enhance the consistency between the saliency map and the picked proposal, i.e.,

$$\ell_{psl}(I_n, O_n; \mathbf{w}) = \frac{1}{W \times H} \|S_n - O_n\|^2, \tag{7}$$

$$\text{where } O_n = \arg\min_{O \in \mathcal{O}_n} \|S_n - O\|^2. \tag{8}$$

In Eq. (8), $\mathcal{O}_n$ is the pool of object proposals produced for I_n by using the adopted proposal generation algorithm. We pick the proposal $O_n \in \{0, 1\}^{W \times H}$ which best matches the saliency map S_n. The saliency map S_n is optimized to be consistent with O_n via Eq. (7). The idea behind this proposal loss is intuitive: The object proposal covering the discriminative parts, i.e., consistent with S_n, likely covers the non-discriminative parts at the same time. This property is leveraged to enforce the generator $g(\cdot)$ to highlight the non-discriminative parts along with the discovered discriminant parts. Consequently, this loss reduces false negatives, and facilitates the detection of complete salient objects. In ℓ_{seg}, the pseudo ground truth is yielded by picking superpixels individually. It does not necessarily maintain the whole objects. In contrast, the goal of an object proposal algorithm is to generate at least one proposal that can cover the whole object, and we can pick the top-ranked proposal via Eq. (8) to overcome issues of incomplete objects or objects with holes.

C. Optimization process

The objective in Eq. (1) is differentiable and convex, and can be efficiently and effectively optimized with *stochastic gradient descent* (SGD). An iterative method is adopted to sequentially update superpixel masks $\{G_n\}$, object proposals $\{O_n\}$, and CNN parameters $\mathbf{w}$. The extraction of superpixels and object proposals is carried out before executing our method. The resultant superpixels and object proposals remain fixed during the iterative process of our method.

When running the proposed method, at each epoch, we first fix the CNN parameters $\mathbf{w}$ and apply the generator $g(\cdot)$ to the training images to get the saliency maps, $\{S_n\}$. Then, we refer to the generated saliency maps and pick the superpixels to produce the masks $\{G_n\}$ via Eq. (6). The most consistent proposals $\{O_n\}$ are selected based on the generated saliency maps via Eq. (8). The generated masks $\{G_n\}$ and the selected object proposals $\{O_n\}$ serve as the pseudo ground truth for optimizing the generator based on the objective function in Eq. (1). The same steps are repeated for each epoch. The optimization is finished until convergence or reaching the maximum epoch number. Algorithm 1 summarizes the optimization procedure.

It is worth mentioning that the superpixels and object proposals are only adopted in the training stage. During testing, the saliency map of a test image is obtained by applying the learned generator to the test image.

D. Implementation details

We implemented the proposed network based on MatConvNet [54]. ResNet-50 [55] is adopted as the image-level classifier $f(\cdot)$, because using other network architectures, such as AlexNet [14] or VGG-16/19 [56], sometimes results in

the vanishing gradient problem. The two-class classifier $f(\cdot)$ is pre-trained on ImageNet [57] 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 map generator is developed based on the VGG-16 [56] setting of FCN [53] with the same batch size, weight decay, and momentum except for the last layer. We replace the activation function *softmax* in the last layer with the *sigmoid* function. The output of the sigmoid function is the estimated saliency map. The learning rate is set to 0.00001, and fixed during training. The maximum number of epochs is set to 200. In the first 100 epochs, we optimize Eq. (1) with loss functions ℓ_{seg} and ℓ_{psl} removed because the initial model is not stable enough to generate reliable superpixel masks and select plausible object proposals. Superpixel masks and object proposals of low quality will drop the performance. In the last 100 epochs, the four loss functions are jointly optimized. Data augmentation including vertical flip, horizontal flip, and rotation at $90, 180, 270$ degrees, is used to avoid over-fitting. In addition, because the classifier $f(\cdot)$ requires the inputs of the same size, each training image is resized to the resolution 384×384 in advance.

For the set of superpixels M_n used in the segmentation loss Eq. (4), the superpixel extraction algorithm SLIC [58] implemented in VLFeat [59] is adopted to decompose an image into superpixels because of its computational efficiency, better compactness and regularity. The average number of superpixels in an image is about 361. For generating the pool of object proposals $\mathcal{O}_n$ used in the proposal loss Eq. (8), we use the fast object proposal generation algorithm, *geodesic object proposal* (GOP) [60]. According to the weakly supervised setting of this work, the unsupervised setting of GOP is adopted. The number of the generated proposals for an image ranges between 200 and 1100. The parameters of SLIC and GOP are the same as those in their demo codes for superpixel extraction and unsupervised proposal generation, respectively.

Algorithm 1 The Optimization Procedure

Input: Object image set: D_{obj}; Background image set: D_{bg}; Maximum number of epochs: T;

Train the image classifier $f(\cdot)$; (Sec. III-D)
Extract the superpixels for each image; (Sec. III-D)
Compute the object proposals for each image; (Sec. III-D)
Initialize the map generator $g(\cdot)$; (Sec. III-D)
for Epoch: 1, ..., T **do**
    Generate saliency maps $\{S_n\}$ with $g(\cdot)$, $\forall I_n \in D_{obj}$;
    Update $\{G_n\}$ with $\{S_n\}$ via Eq. (6), $\forall I_n \in D_{obj}$;
    Update $\{O_n\}$ with $\{S_n\}$ via Eq. (8), $\forall I_n \in D_{obj}$;
    Optimize objective in Eq. (1) with $\{G_n\}$ and $\{O_n\}$;
    if convergence **then**
        Return $g(\cdot)$;
    end if
end for

Output: Saliency map generator $g(\cdot)$;

IV. Experimental results

This section evaluates the proposed approach. We first describe the datasets and the metrics for performance evaluation. Next, we report the sensitivity analysis on the model parameters and assess the impacts of each loss function. Finally, we compare the proposed approach with the stare-of-the-art weakly supervised and fully supervised approaches. These methods are compared both quantitatively and visually.

A. Datasets and evaluation criterion

We evaluated our proposed method on three benchmarks for top-down saliency detection, including Graz-02 [15], PASCAL VOC-07, and VOC-12 [16]. We chose the three datasets because they are composed of real-world images with large intra-class variations, occlusions and background clutters. They have been widely used in the literature of top-down saliency detection, such as [6]–[10], [37].

1) Graz-02: The Graz-02 dataset [15] contains images of three object categories, bike, car and person, and a background category. Each category has 300 images of resolution 640×480. The ground truth in the form of pixel-level object masks are provided for the performance evaluation. Following the setting used in previous papers [6], [8], [9], the odd numbered 150 images from each category served as the training data, while the rest were treated as the test data. Three saliency models were trained, one for each object category.

2) PASCAL VOC-07 and VOC-12: The PASCAL VOC-07 and VOC-12 datasets are more challenging and difficult than the Graz-02 dataset because more variations, occlusions and background clutters are present in the images. The PASCAL VOC-12 [16] dataset consists of 20 object categories. It contains $5{,}717$ training images and $5{,}823$ validation images in the tasks of object classification and detection, while it has $1{,}464$ training images and $1{,}449$ validation images in the segmentation task. For all the three tasks, the ground truth of the test images are not available. Following the evaluation protocols adopted in previous work [10], [37], [43], we used the $5{,}717$ training images in the classification task as the training data, while adopting the $1{,}449$ validation images, which have pixel-wise object masks, in the segmentation task as the testing data. For each object category, only images where the target object are present were used for evaluation.

PASCAL VOC-07 is a subset of PASCAL VOC-12, but the ground truth of the 210 test images for segmentation is provided. For PASCAL VOC-07, because CNNs require a lot of training images, the same training images were used to train the models, and the 210 test images were used for testing. Following the setting used in previous work [6], [8], [9], all models were evaluated on the 210 test images no matter whether the target objects are present or not.

3) Evaluation criterion: The *precision rate at equal error rate* (Prec@EER), was adopted to measure the performance. Following the previous researches [6]–[10], [37], the saliency maps in the Graz-02 and PASCAL VOC-07 datasets were not binarized when computing Prec@EER. For the PASCAL VOC-12 dataset, we used the same setting as He et al. [10]

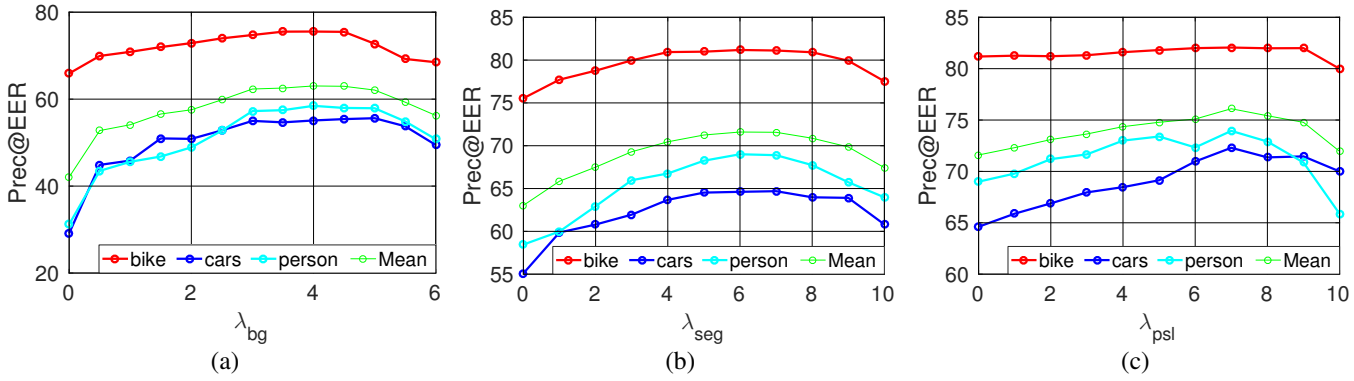

Fig. 4. The performances of our approach in Prec@EER with different vales of weighting parameter (a) λ_{bg}, (b) λ_{seg}, and (c) λ_{psl} on the Graz-02 dataset, when adding the three loss functions ℓ_{bg}, ℓ_{seg}, and ℓ_{psl} one by one in the order.

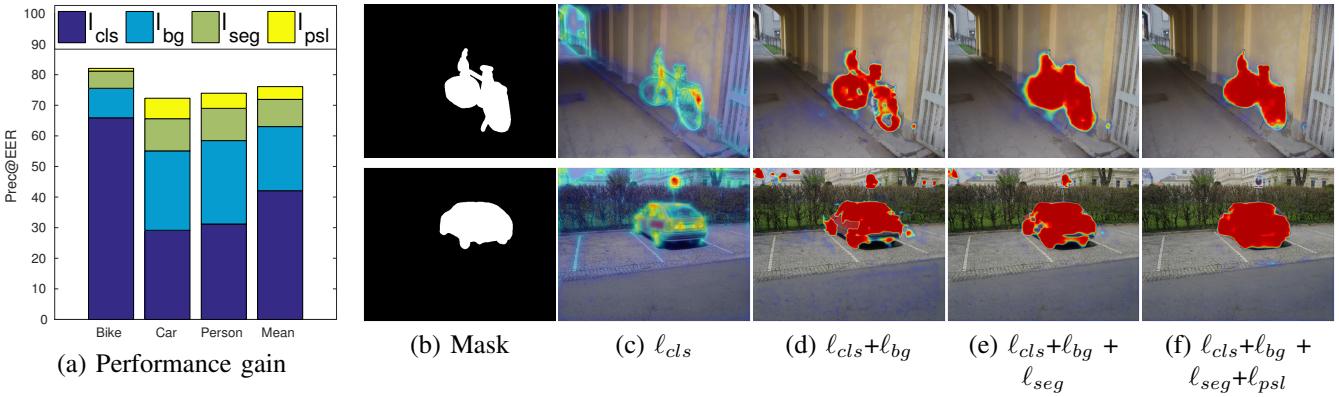

Fig. 5. (a) The performance gains in Prec@EER obtained by adding the four loss functions, i.e., ℓ_{cls}, ℓ_{bg}, ℓ_{seg}, and ℓ_{psl}, one by one on the Graz-02 dataset. (b) The ground truth masks for the two examples. (c) $\sim$ (f) The saliency maps produced on two examples when the four loss functions are sequentially added to the objective function.

to evaluate our model. In their work [10], the saliency maps were first binarized with every integer threshold in the range of $[0, 255]$, and then Prec@EER was computed by using the threshold with the smallest difference between precision and recall.

B. Results on the Graz-02 dataset

In the following, we first conduct model analysis to determine the values of the parameters in our method, and then compare our method with the existing methods on the Graz-02 dataset.

1) Model analysis: Following the previous work [6], [9], we analyze our model and empirically select the hyper-parameters on the Graz-02 test data. The proposed objective in Eq. (1) consists of four loss functions. Except the classification loss ℓ_{cls}, the other three loss functions, ℓ_{bg}, ℓ_{seg}, and ℓ_{psl}, are associated with weighting parameters, i.e., λ_{bg}, λ_{seg}, and λ_{psl}, respectively. We conduct sensitivity analysis of the three parameters, and assessed the effect of adopting these loss functions. The classification loss ℓ_{cls} is always included in the objective function with the weight 1. We first add the background loss ℓ_{bg} for removing false positives in saliency maps. Fig. 4(a) reports the performance of the proposed method by varying λ_{bg}. 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 4.

Next, the third loss ℓ_{seg} is included to preserve the object boundaries and remove the noise. The performance of our approach with different values of λ_{seg} is similarly reported in Fig. 4(b). The loss ℓ_{seg} moderately enhances saliency detection. The parameter λ_{seg} is fixed to 6. Finally, the fourth loss ℓ_{psl} is introduced to cover the non-discriminative object parts and highlight the complete objects in the images. As shown in Fig. 4(c), this loss enhances the performance of saliency detection. The parameter λ_{psl} is set to 7. The optimal values of these parameters have similar trends among the three object categories. We fix the parameters, $(\lambda_{bg}, \lambda_{seg}, \lambda_{psl}) = (4, 6, 7)$, for all categories 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_{seg}, and ℓ_{psl}, to the objective function. The results in Fig. 5(a) indicate that each loss function makes its own contribution to saliency detection for all the three object categories. To get insight into the gains, two examples of the detected saliency maps generated through the procedure of sequentially adding the four loss functions are given in Figs. 5(c) $\sim$ 5(f). With only the classification loss ℓ_{cls}, the target objects, *bicycle* and *car*, are detected, but many false alarms occur in Fig. 5(c). From Fig. 5(c) to Fig. 5(d), the background loss ℓ_{bg} is added, and it helps remove most false alarms. It can be observed that the background loss, separating background regions from objects, has objects

TABLE I
THE PERFORMANCES IN PREC@EER (%) OF DIFFERENT APPROACHES, INCLUDING UNSUPERVISED (US), FULLY SUPERVISED (FS), AND WEAKLY SUPERVISED (WS) ONES, ON THE GRAZ-02 DATASET. SP AND SM REPRESENT THE USE OF SUPERPIXELS AND EXISTING SALIENCY MAPS DURING INFERENCE, RESPECTIVELY.

Group	Method	Setting	Bike	Car	Person	Mean
Bottom-up	MB [61]	US	54.7	39.0	52.0	48.6
	MST [62]	US	50.1	38.8	51.3	46.7
	WSS [32]	WS	64.7	71.6	64.0	66.8
	HDCT [25]	FS	55.9	43.8	53.0	50.9
	DRFI [26]	FS	51.3	49.6	59.6	53.5
	Amulet [33]	FS	78.5	75.7	78.4	77.5
	UCF [34]	FS	70.8	70.7	76.2	72.6
	PiCANet [35]	FS	79.7	82.1	85.0	82.3
	C2SNET [36]	FS	79.8	80.9	83.0	81.2
Top-down	ILC [63]	FS	71.9	64.9	58.6	65.1
	SP-Nei. [64]	FS	72.2	72.2	66.1	70.2
	Shape mask [65]	FS	61.8	53.8	44.1	53.2
	Patch-CRF [6]	FS	62.4	60.0	62.0	61.3
	SP-CRF [7]	FS	73.9	68.4	68.2	70.2
	LCCSC [8]	FS	76.2	71.2	64.1	70.5
	R-ScSPM [9]	FS	77.6	71.9	67.0	72.1
	R-ScSPM [9]	WS	67.5	56.5	57.6	60.5
	R-ScSPM+ [11]	WS	-	-	-	69.1
	Ours (prior) [37]	WS	78.9	66.6	64.2	69.9
	Ours	WS	82.1	78.5	75.0	78.5
	R-ScSPM+ [11]	WS+SP+SM	84.1	81.5	81.8	82.5

detected more confidently. From Fig. 5(d) to Fig. 5(e), the added segmentation loss ℓ_{seg} makes the saliency maps much sharper, since this loss helps preserve the object boundaries and remove the noise. From Fig. 5(e) to Fig. 5(f), we find that adopting the loss ℓ_{psl} can identify the non-discriminative object parts, highlight the complete objects, and also further remove the noise. In the *car* image of Fig. 5(e), the detection result is incomplete since some holes are present inside the car. The unfavorable effect results from picking superpixels individually. In Fig. 5(f) where the loss regarding object proposals has been incorporated, it is obvious that the object can be detected more completely.

2) Comparison with the state-of-the-art methods: For the Graz-02 dataset, we compare our proposed method with the state-of-the-art methods, and report their performances in TABLE I, where the field *setting* denotes the supervision condition of training data, including *unsupervised* (US), *weakly supervised* (WS), and *fully supervised* (FS) settings.

In TABLE I, the bottom-up methods [25], [26], [61], [62] on Graz-02 identify salient objects without using any prior information of the target category. Despite the broad applicability, they do not perform very well for category-specific saliency detection. The CNN-based methods [33]–[36] for supervised bottom-up saliency detection use large-scale training data with annotated object masks, so they often outperform the conventional bottom-up methods. However, our method can still achieve the better or comparable performance without using training data with annotated masks. Compared to WSS [32] which also uses image-level labeled training data, our method reaches the much better performance. The main reason is that WSS [32] is a bottom-up method while our method handles top-down saliency detection and addresses salient objects of a target category.

TABLE II
THE AVERAGE RUN TIME (IN SECONDS) OF THE COMPETING METHODS AND OUR METHOD ON THE GRAZ-02 DATASET.

Method	MB [61]	MST [62]	Patch-CRF [6]	SP-CRF [7]
Time (Sec)	0.0263	0.1142	3.0940	30.2928
Speedup	1151.8×	265.3×	9.8×	1×
Method	Examplar [10]	Ours (prior) [37]	Exc. BP [51]	Ours
Time (Sec)	2.1470	5.2950	0.0632	0.0151
Speedup	14.1×	5.7×	479.3 ×	2006.1×

Instead, fully supervised, top-down methods [6]–[8], [63]–[65] 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 [9] 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 [9] and our prior work [37] by large margins around 18% and 8.6% in Prec@EER, respectively. The large performance gain of our method over its prior work [37] reveals that the newly introduced segmentation- and object-proposal-based losses compensate for the lack of pixel-wise annotated training data in the weakly supervised setting. R-ScSPM+ [11] outperforms the proposed method because it uses two-step post-processing, namely bottom-up saliency map fusion and multi-scale superpixel-averaging. Under the same setting where post-processing is turned off, our method outperforms R-ScSPM+ [11] by a large margin. It is also worth mentioning that our method even achieves a remarkably better performance than the state-of-the-art fully supervised methods. Thus, we believe that the proposed losses could also benefit the supervised setting and likely advance the methods in this category.

To gain insight into the quantitative results, Fig. 6 shows some detected saliency maps by different approaches. The bottom-up approaches, MB [61] and DRFI [26], tend to misclassify non-target objects as the salient regions. These false positives are caused due to the lack of category-specific information in training data, and are prone to occur in the regions of high contrast, such as windows, bags, and clothes. Compared to MB and DRFI, the top-down methods, Patch-CRF [6] and SP-CRF [7], can yield more 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 positives. Second, their features are extracted from a patch [6] or a superpixel [7] to reduce the complexity. The resultant feature maps cannot preserve the fine structures in the images very well, and may have the unfavorable block effect.

In our prior work [37], though a postprocessing for enforcing spatial coherence is employed to make the generated saliency maps smooth and remove the false negatives, it sometimes results in over-smooth saliency maps. In the proposed method, the postprocessing is replaced with the segmentation and object proposal information. As can be seen in Fig. 6, our method does not suffer from the aforementioned issues. It can

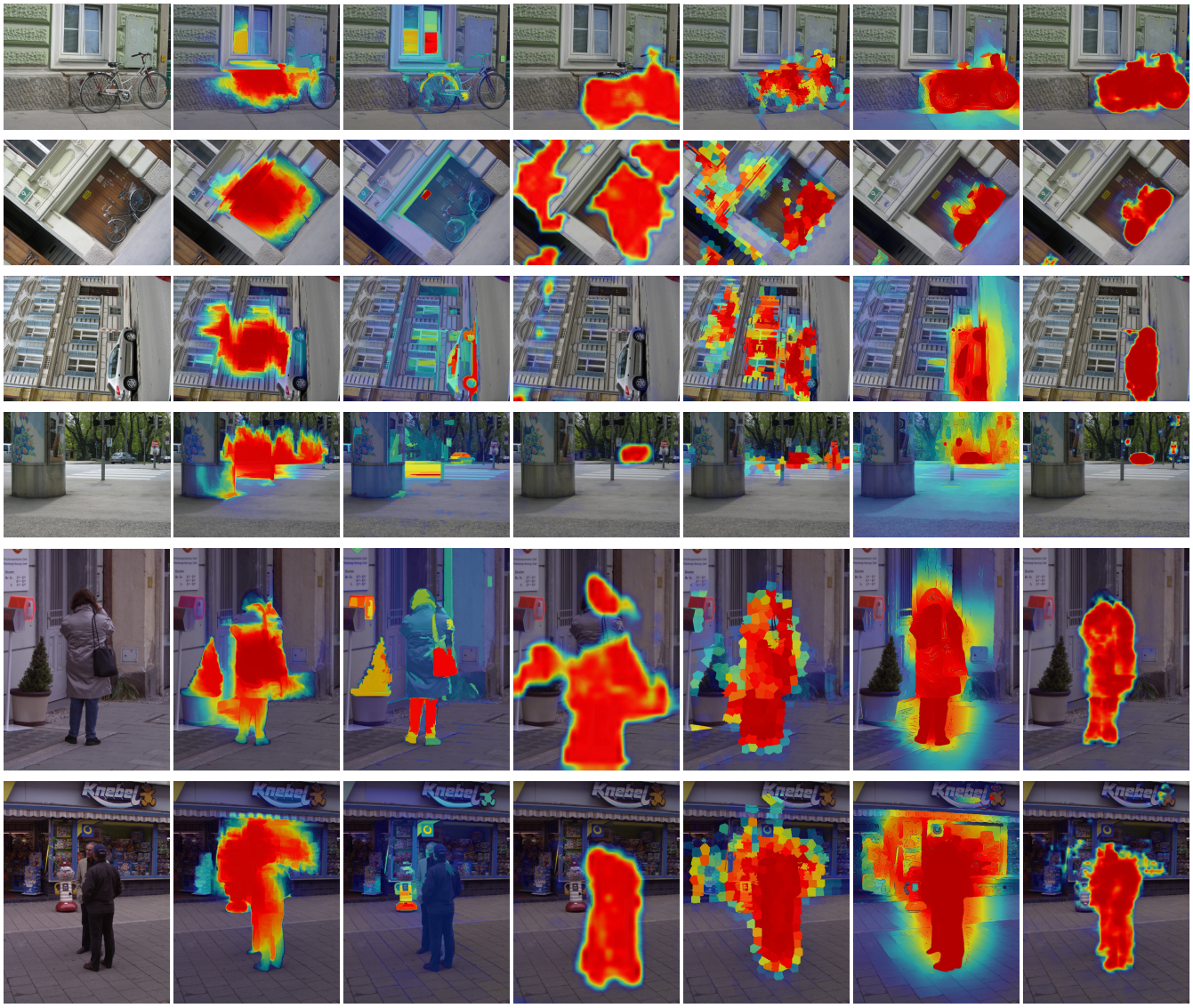

Fig. 6. The saliency maps detected by our approach and the competing approaches on the Graz-02 dataset. In the six examples (rows), the target categories are *bike* in the first two rows, *car* in the middle two rows, and *person* in the last two rows.

better preserve the object boundaries than the prior work [37] and produce saliency maps of higher quality.

By replacing the time-consuming postprocessing with the integrated CNN training with the proposed segmentation and object proposal losses, the proposed method outperforms the competing methods not only on results but also on the running time. We compare the running time of different methods, including the real-time bottom-up methods [61], [62], non CNN-based top-down methods [6], [7], CNN-based top-down methods [10], [37], and top-down neural attention [51]. Note that extracting superpixels and object proposals is not required for our method during inference. TABLE II reports the average running time of these competing methods and our method for predicting the saliency map of an image in the Graz-02 dataset. He et al. [10] only released the test code trained on PASCAL VOC-12, so the performance of their method on the Graz-02 dataset is unknown and is not reported in TABLE I. Nevertheless, we compare our method with their method in terms of running time, as shown in TABLE II. The main computation of our method lies in executing the map generator, FCN. Note that we conducted the experiments on images of resolution 384×384 on NVIDIA GTX Titan. The lower image resolution and the faster GPU card make our running time less than that of FCN reported in [53].

The proposed method is faster because of some nice properties. First, our method employs a CNN model, and doesn't require to extract potentially costly hand-crafted features from images. For example, it is faster than the methods [6], [7] becasue they spend lots of computation on extracting SIFT or objectness scores. Second, compared with other CNN-based methods, our method performs saliency detection with just one forward pass, namely applying the learned map generator to an input image. In contrast, the sliding window method [10] requires multiple forward passes and extra computation for

TABLE III
PREC@EER (%) ON PASCAL VOC-07.

Method	Setting	Avg.	A.P.	Bike.	Bird	Boat	Bottle.	Bus	Car	Cat	Chair	Cow	D.T.	Dog	Horse	M.B.	P.S.	P.P.	Sheep	Sofa	Train	TV
Patch-CRF [6]	FS	16.2	15.2	39.0	9.4	5.7	3.4	22.0	30.5	15.8	5.7	8.0	11.1	12.8	10.9	23.7	42.0	2.0	20.2	10.4	24.7	10.5
LCCSC [8]	FS	23.4	13.3	33.2	22.1	11.2	8.6	33.5	37.2	14.3	3.9	22.3	23.0	14.9	25.0	30.6	38.9	16.4	36.3	18.3	29.2	36.3
R-ScSPM [9]	WS	18.6	41.0	19.5	9.9	10.2	1.5	27.3	34.0	14.7	14.1	21.2	9.9	7.5	14.8	30.9	36.4	8.8	18.5	7.1	31.5	13.6
Ours (prior) [37]	WS	23.5	28.3	23.7	51.7	7.8	0.0	18.5	39.1	33.7	1.4	18.3	11.6	24.7	24.7	35.0	62.3	11.4	35.8	2.3	11.8	28.3
Ours	WS	27.5	28.8	32.2	59.2	11.0	0.0	31.0	45.1	46.9	0.6	23.2	19.5	21.6	34.0	49.2	45.0	22.3	30.0	1.4	22.4	25.7
SP-CRF [7]	FS*	41.9	49.4	46.6	33.7	60.9	26.1	51.8	35.1	64.9	21.1	34.8	43.7	35.1	41.4	71.4	32.6	42.0	42.5	13.8	63.8	27.8
Ours	WS*	47.2	54.2	54.9	67.7	17.6	0.0	68.0	57.8	90.0	10.7	38.0	38.7	64.1	63.4	81.4	20.9	29.4	77.5	10.8	63.2	35.6

TABLE IV
PREC@EER (%) ON PASCAL VOC-12. * INDICATES BOTTOM-UP SALIENCY METHODS. SP AND SM REPRESENT THE USE OF SUPERPIXELS AND EXISTING SALIENCY MAPS DURING INFERENCE, RESPECTIVELY.

Method	Setting	Avg.	A.P.	Bike.	Bird	Boat	Bottle.	Bus	Car	Cat	Chair	Cow	D.T.	Dog	Horse	M.B.	P.S.	P.P.	Sheep	Sofa	Train	TV
WSS [32]	WS*	51.2	65.6	40.0	51.0	48.2	35.5	69.4	45.0	69.0	25.4	69.1	35.0	66.1	69.5	65.2	43.3	24.7	68.6	35.8	68.7	29.4
Amulet [33]	FS*	60.4	87.0	34.8	69.0	58.0	36.1	85.7	52.8	78.7	22.3	83.9	35.1	78.5	80.8	71.7	57.8	25.2	83.8	43.2	82.0	42.2
UCF [34]	FS*	58.7	83.0	35.1	69.2	58.4	38.2	80.3	52.6	74.0	23.1	83.7	33.3	76.3	78.3	70.1	56.2	26.5	78.9	43.9	79.0	34.9
PiCANet [35]	FS*	62.9	81.2	40.8	73.4	69.1	39.9	86.8	55.1	81.4	22.5	86.3	35.9	80.8	81.0	73.4	60.9	22.6	89.3	48.3	81.4	48.6
C2SNET [36]	FS*	62.7	84.8	37.3	73.6	69.1	39.6	86.0	53.7	81.5	25.6	83.2	36.5	80.6	82.4	74.1	56.2	26.6	87.0	47.6	82.6	45.6
R-ScSPM+ [11]	WS+SP+SM	61.4	71.2	22.3	74.9	39.9	52.5	82.7	58.9	83.4	27.1	81.1	49.3	82.4	77.9	74.2	69.8	31.9	81.4	49.8	63.2	53.3
Patch-CRF [6]	FS	15.6	14.7	28.1	9.8	6.1	2.2	24.1	30.2	17.3	6.2	7.6	10.3	11.5	12.5	24.1	36.7	2.2	20.4	12.3	26.1	10.2
SP-CRF [7]	FS	40.4	46.5	45.0	33.1	60.2	25.8	48.4	31.4	64.4	19.8	32.2	44.7	30.1	41.8	72.1	33.0	40.5	38.6	12.2	64.6	23.6
Examplar [10]	FS	56.2	55.9	37.9	45.6	43.8	47.3	83.6	57.8	69.4	22.7	68.5	37.1	72.8	63.7	69.0	57.5	43.9	66.6	38.3	75.1	56.7
GMP [43]	WS	48.1	48.9	42.9	37.9	47.1	31.4	68.4	39.9	66.2	27.2	54.0	38.3	48.5	56.5	70.1	43.2	42.6	52.2	34.8	68.1	43.4
Exc. BP [51]	WS	45.3	50.7	32.5	48.4	30.2	36.8	59.3	36.6	54.4	21.6	57.6	40.4	59.0	47.5	61.4	48.4	28.7	57.5	35.8	48.7	51.5
R-ScSPM+ [11]	WS	50.2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Ours (prior) [37]	WS	50.0	64.5	46.7	50.2	29.6	0.0	75.3	60.1	73.4	16.0	39.5	40.9	81.8	59.9	72.5	72.0	37.6	58.9	45.3	43.5	32.9
Ours	WS	56.8	71.6	47.7	64.6	32.4	0.0	77.8	69.4	81.9	19.5	48.9	39.9	76.8	71.3	75.0	87.4	42.0	75.8	67.8	54.0	32.1

the gradients via back-propagation [51]. Third, our method does not need any optimization or postprocessing process in the test stage. On the contrary, our prior work [37] computes the edge probability to enhance map smoothness and preserve object boundaries in saliency detection, greatly degrading the efficiency. Like other CNN-based methods, our method can be dramatically accelerated by GPU parallel computing. Thus, the running time of our method is even less than that of the real-time bottom-up methods [61], [62].

C. Results on the PASCAL VOC-07 and VOC-12 datasets

In the following, we compare our method with the state-of-the-art methods on the PASCAL VOC-07 and PASCAL VOC-12 datasets. The same procedure as that in Graz-02 is adopted for tuning the parameters. The parameter values are set and fixed for each dataset. The performances of different approaches on PASCAL VOC-07 and VOC-12 are reported in TABLE III and TABLE IV, respectively. In both tables, the supervision condition of training data, the average performance, and the performance on each category are given for each method.

We first discuss the results on PASCAL VOC-12. The competing methods include five CNN-based bottom-up saliency detection methods [32]–[36], two top-down saliency detection methods [6], [7], the state-of-the-art method [10], a method based on object localization [43], a method based on neural attention [51], and our prior work [37]. The competing methods [10], [32], [37], [43], [51] are also based on CNNs. The competing methods [6], [7], [10], [33]–[36] adopt the fully supervised (FS) setting, while the others [32], [37], [43], [51] adopt the weakly supervised (WS) one.

In TABLE IV, our proposed method performs favorably against all competing methods. The WS localization methods [43], [51] aim at object localization, and often detect merely the discriminative object parts rather the whole objects. Our method instead can identify the full extent of the target objects and preserve the object boundaries. The performance gains of our method over the two methods [43], [51] are significant, around 8.7% and 11.5% respectively. The WS bottom-up saliency method [32] detects only the most salient objects instead of all salient objects, so it performs worse especially when multiple salient objects are present. The performance gain of using our method, about 5.6%, is significant. Owing to post-processing, R-ScSPM+ [11] achieves better performance than our method. Nevertheless, under the setting where no post-processing steps are used, our method can outperform R-ScSPM+ [11]. Although adopting the weakly supervised setting, our method even achieves a slightly better performance than the state-of-the-art FS method [10]. The encouraging result implies that integrating the information of segmentation and object proposals into learning the two CNN modules in our method can compensate for the lack of the fully labeled training data. Our method falls behind the fully supervised bottom-up methods [33]–[36], but these fully supervised methods require training data with annotated object masks. Instead, our method uses training data with image-level labels, and thus the annotation cost is greatly reduced.

On the PASCAL VOC-07 dataset, we compare our method with the state of the arts, including the FS methods [6]–[8] and the WS methods [9], [37]. The results are shown in TABLE III. Note that our method is evaluated with two different experimental settings, one for the comparison with the method in [7] and the other for other methods: * in field *setting* of

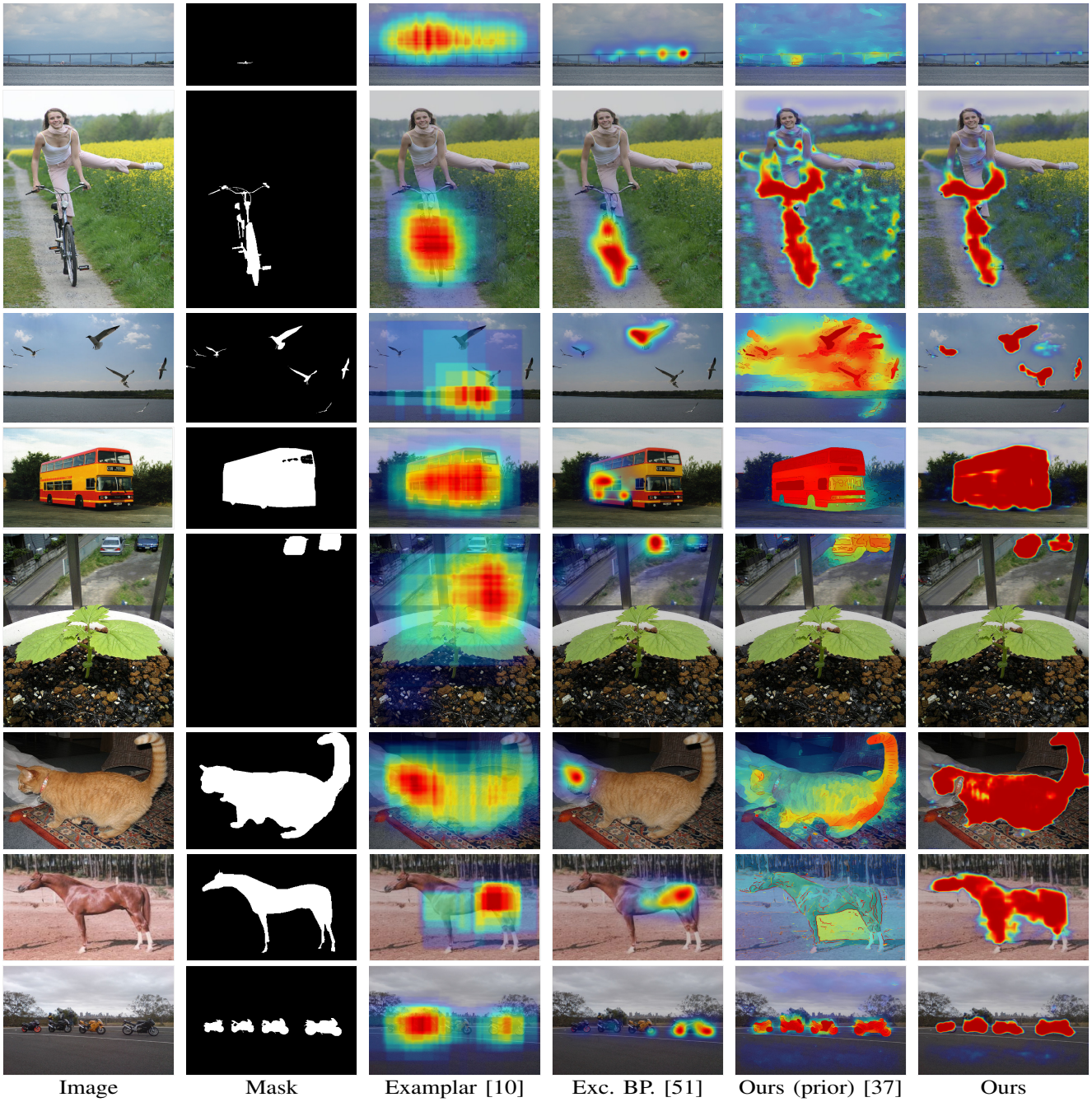

Fig. 7. The saliency maps detected by different approaches on the PASCAL VOC-12 dataset. For top to bottom, the target object categories are *airplane, bicycle, bird, bus, car, cat, horse,* and *motorbike*, respectively.

TABLE III indicates the former setting where the zero-valued saliency maps are manually assigned to images where no target object is present. In both settings, our method outperforms all other methods. The results demonstrate the effectiveness of our method. In TABLE III and TABLE IV, our method provides significant gains over our prior work [37], because it further considers two reliable cues, i.e., the segmentation-based loss and the proposal-based loss. The former helps preserve the object boundaries and exclude noise, while the latter can discover more non-discriminative object parts and reduce false negatives. As we will show in the following, the two visual cues help generate saliency maps of higher quality, and are essential to the performance improvement.

Fig. 7 shows the detected saliency maps on PASCAL VOC-12 for visually comparing different approaches to saliency detection. It can be observed that there are some limitations of the FS top-down method [10] and the WS top-down method [51]. First, the saliency maps generated by the two approaches are too coarse to preserve the object boundaries. Second, only the object parts rather than the whole objects are discovered. This phenomenon is evident in the third, fourth, sixth, seventh and eighth examples (rows). Third, when there exist more than one object, the two approaches sometimes fail to detect all the objects, e.g., the results in the third,

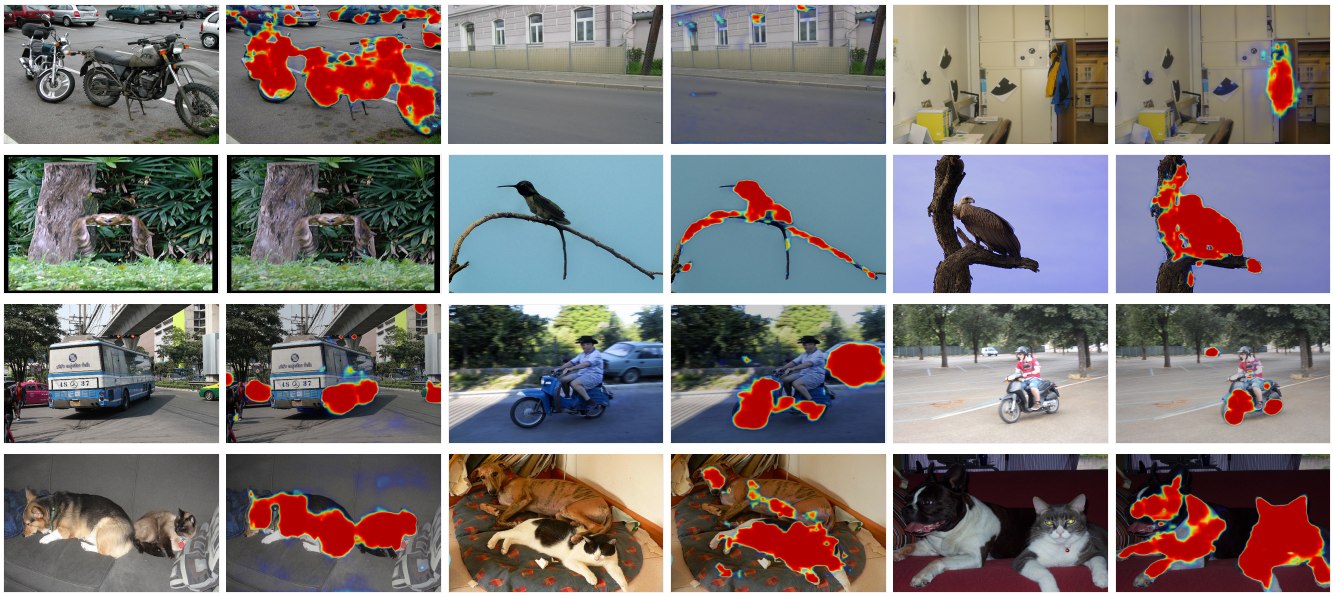

Fig. 8. Some failure cases of our approach. The cases in the first row come from Graz-02, and the rest come from PASCAL VOC-12. In the first row, the first, third, and fifth images are the background images from the categories *bike, car*, and *person*, respectively, and their corresponding saliency maps are shown in the second, fourth and sixth images, respectively. The proposed method generates false positives because the background contains similar features to the target object. In the last three rows, the target objects are from the categories *bird, car,* and *cat*, respectively. Again, errors occur due to similar features between the object and the background.

fifth and eighth rows. Although the FS method [10] achieves the performance comparable to ours, it tends to produce coarse saliency maps and can't preserve object boundaries. Compared to our prior work [37], the proposed method gives less false positives and better boundary preservation, which can be attributed to the newly added segmentation-based loss function. In addition, our method can pick and leverage object proposals so that the whole salient objects are highlighted more sharply and uniformly. It is also worth mentioning that our method can perform well in the challenging cases such as small objects in the first row, multiple objects in the third, fifth, and eighth rows, objects of complex shapes in the second row, and objects with large intra-object variations in the fourth row.

D. Failure cases

We show some failure cases of our approach in Fig. 8. Most failure cases are caused by the high similarity between target objects and the background, including objects of non-target categories. In the first row, 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. In the second row, the high similarity between target objects (birds) and background (trees) causes the false negatives in the first example and the false positives in the last two examples. In the third row, common object parts shared across categories, i.e., the tires of buses, cars, and motorbikes, result in false positives. In the last row, multiple object categories having similar appearance, namely cats and dogs here, lead to false alarms.

V. 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 CNN modules, i.e., 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, superpixel-based smoothing, and object-like proposal selection, with which the unfavorable 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 on three benchmarks for saliency detection, including the Graz-02, PASCAL VOC-07, and PASCAL VOC-12 datasets, demonstrate that our method remarkably outperforms the existing weakly supervised methods and even achieves better results than the state-of-the-art fully supervised methods. In the 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.

References

[1] L. Zhu, Z. Chen, X. Chen, and N. Liao, "Saliency & structure preserving multi-operator image retargeting," in *Proc. Int'l Conf. Acoustics, Speech and Signal Processing*, 2016.

[2] S. Hong, T. You, S. Kwak, and B. Han, "Online tracking by learning discriminative saliency map with convolutional neural network," in *Proc. Int'l Conf. Machine Learning*, 2015.

[3] W. Wang, J. Shen, and F. Porikli, "Saliency-aware geodesic video object segmentation," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2015.

[4] 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 *Proc. Conf. Computer Vision and Pattern Recognition*, 2011.

[5] Z. Ren, S. Gao, L.-T. Chia, and I. W.-H. Tsang., "Region-based saliency detection and its application in object recognition," *IEEE Trans. on Circuits and Systems for Video Technology*, 2014.

[6] J. Yang and M.-H. Yang, "Top-down visual saliency via joint CRF and dictionary learning," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2012.

[7] A. Kocak, K. Cizmeciler, A. Erdem, and E. Erdem, "Top-down saliency estimation via superpixel-based discriminative dictionaries," in *Proc. British Conf. Machine Vision*, 2014.

[8] H. Cholakkal, D. Rajan, and J. Johnson, "Top-down saliency with locality-constrained contextual sparse coding," in *Proc. British Conf. Machine Vision*, 2015.

[9] H. Cholakkal, J. Johnson, and D. Rajan, "Backtracking ScSPM image classifier for weakly supervised top-down saliency," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2016.

[10] S. He, R. Lau, and Q. Yang, "Exemplar-driven top-down saliency detection via deep association," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2016.

[11] H. Cholakkal, J. Johnson, and D. Rajan, "Backtracking spatial pyramid pooling-based image classifier for weakly supervised top–down salient object detection," *IEEE Trans. on Image Processing*, 2018.

[12] K.-J. Hsu, Y.-Y. Lin, and Y.-Y. Chuang, "Augmented multiple instance regression for inferring object contours in bounding boxes," *IEEE Trans. on Image Processing*, 2014.

[13] F.-J. Chang, Y.-Y. Lin, and K.-J. Hsu, "Multiple structured-instance learning for semantic segmentation with uncertain training data," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2014.

[14] A. Krizhevsky, I. Sutskever, and G. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, 2012.

[15] A. Opelt, A. Pinz, M. Fussenegger, and P. Auer, "Generic object recognition with boosting," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 2006.

[16] M. Everingham, L. V. Gool, C. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (VOC) challenge," *Int. J. Computer Vision*, 2010.

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

[18] Y.-F. Ma and H.-J. Zhang, "Contrast-based image attention analysis by using fuzzy growing," in *Proc. ACM Conf. Multimedia*, 2003.

[19] F. Liu and M. Gleicher, "Region enhanced scale-invariant saliency detection," in *Proc. Conf. Multimedia and Expo*, 2006.

[20] M.-M. Cheng, N. J. Mitra, X. Huang, P. H. S. Torr, and S.-M. Hu, "Global contrast based salient region detection," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 2014.

[21] H. Jiang, J. Wang, Z. Yuan, T. Liu, and N. Zheng, "Automatic salient object segmentation based on context and shape prior," in *Proc. British Conf. Machine Vision*, 2011.

[22] P. Jiang, H. Ling, J. Yu, and J. Peng, "Salient region detection by UFO: Uniqueness, focusness and objectness," in *Proc. Int'l Conf. Computer Vision*, 2013.

[23] K.-Y. Chang, T.-L. Liu, H.-T. Chen, and S.-H. Lai, "Fusing generic objectness and visual saliency for salient object detection," in *Proc. Int'l Conf. Computer Vision*, 2011.

[24] Y. Jia and M. Han, "Category-independent object-level saliency detection," in *Proc. Int'l Conf. Computer Vision*, 2013.

[25] J. Kim, D. Han, Y.-W. Tai, and J. Kim, "Salient region detection via high-dimensional color transform," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2014.

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

[27] S. Li, H. Lu, Z. Lin, X. Shen, and B. Price, "Adaptive metric learning for saliency detection," *IEEE Trans. on Image Processing*, 2015.

[28] L. Wang, H. Lu, X. Ruan, and M.-H. Yang, "Deep networks for saliency detection via local estimation and global search," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2015.

[29] L. Wang, L. Wang, H. Lu, P. Zhang, and X. Ruan, "Saliency detection with recurrent fully convolutional networks," in *Proc. Euro. Conf. Computer Vision*, 2016.

[30] S. He, R. Lau, W. Liu, Z. Huang, and Q. Yang, "SuperCNN: A superpixelwise convolutional neural network for salient object detection," *Int. J. Computer Vision*, 2015.

[31] X. Li, L. Zhao, L. Wei, M.-H. Yang, F. Wu, Y. Zhuang, H. Ling, and J. Wang, "DeepSaliency: Multi-task deep neural network model for salient object detection," *IEEE Trans. on Image Processing*, 2016.

[32] L. Wang, H. Lu, Y. Wang, M. Feng, D. Wang, B. Yin, and X. Ruan, "Learning to detect salient objects with image-level supervision," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2017.

[33] P. Zhang, D. Wang, H. Lu, H. Wang, and X. Ruan, "Amulet: Aggregating multi-level convolutional features for salient object detection," in *Proc. Int'l Conf. Computer Vision*, 2017.

[34] P. Zhang, D. Wang, H. Lu, H. Wang, and B. Yin, "Learning uncertain convolutional features for accurate saliency detection," in *Proc. Int'l Conf. Computer Vision*, 2017.

[35] N. Liu, J. Han, and M.-H. Yang, "PiCANet: Learning pixel-wise contextual attention for saliency detection," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2018.

[36] X. Li, F. Yang, H. Cheng, W. Liu, and D. Shen, "Contour knowledge transfer for salient object detection," in *Proc. Euro. Conf. Computer Vision*, 2018.

[37] K.-J. Hsu, Y.-Y. Lin, and Y.-Y. Chuang, "Weakly supervised saliency detection with a category-driven map generator," in *Proc. British Conf. Machine Vision*, 2017.

[38] D. Li, J.-B. Huang, Y. Li, S. Wang, and M.-H. Yang, "Weakly supervised object localization with progressive domain adaptation," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2016.

[39] H. Bilen and A. Vedaldi, "Weakly supervised deep detection networks," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2016.

[40] V. Kantorov, M. Oquab, M. Cho, and I. Laptev, "Contextlocnet: Context-aware deep network models for weakly supervised localization," in *Proc. Euro. Conf. Computer Vision*, 2016.

[41] A. Bearman, O. Russakovsky, V. Ferrari, and F.-F. Li, "What's the point: Semantic segmentation with point supervision," in *Proc. Euro. Conf. Computer Vision*, 2016.

[42] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2016.

[43] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Is object localization for free? – Weakly-supervised learning with convolutional neural networks," in *Proc. Int'l Conf. Computer Vision*, 2015.

[44] A. Kolesnikov and C. Lampert, "Seed, expand and constrain: Three principles for weakly-supervised image segmentation." in *Proc. Euro. Conf. Computer Vision*, 2016.

[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," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 2016.

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

[47] W. Shimoda and K. Yanai, "Distinct class-specific saliency maps for weakly supervised semantic segmentation," in *Proc. Euro. Conf. Computer Vision*, 2016.

[48] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," in *Proc. Int'l Conf. Learning Representations Worhshop*, 2014.

[49] M. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proc. Euro. Conf. Computer Vision*, 2014.

[50] C. Cao, X. Liu, Y. Yang, Y. Yu, J. Wang, Z. Wang, Y. Huang, L. Wang, C. Huang, W. Xu, D. Ramanan, and T. Huang, "Look and think twice: Capturing top-down visual attention with feedback convolutional neural networks," in *Proc. Int'l Conf. Computer Vision*, 2015.

[51] J. Zhang, Z. Lin, J. Brandt, X. Shen, and S. Sclarff, "Top-down neural attention by excitation backprop," in *Proc. Euro. Conf. Computer Vision*, 2016.

[52] R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. Int'l Conf. Computer Vision*, 2017.

[53] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional models for semantic segmentation," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2015.

[54] A. Vedaldi and K. Lenc, "MatConvNet – convolutional neural networks for matlab," in *Proc. ACM Conf. Multimedia*, 2015.

[55] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2016.

[56] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int'l Conf. Learning Representations*, 2015.
[57] 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 *Proc. Conf. Computer Vision and Pattern Recognition*, 2009.
[58] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "SLIC superpixels compared to state-of-the-art superpixel methods," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 2012.
[59] A. Vedaldi and B. Fulkerson, "VLFeat - an open and portable library of computer vision algorithms," in *Proc. ACM Conf. Multimedia*, 2010.
[60] P. Krähenbühl and V. Koltun, "Geodesic object proposals," in *Proc. Euro. Conf. Computer Vision*, 2014.
[61] J. Zhang, S. Sclaroff, Z. Lin, X. Shen, B. Price, and R. Měch, "Minimum barrier salient object detection at 80 fps," in *Proc. Int'l Conf. Computer Vision*, 2015.
[62] W.-C. Tu, S. He, Q. Yang, and S.-Y. Chien, "Real-time salient object detection with a minimum spanning tree," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2016.
[63] D. Aldavert, A. Ramisa, R. L. de Mantaras, and R. Toledo, "Fast and robust object segmentation with the integral linear classiifier," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2010.
[64] B. Fulkerson, A. Vedaldi, , and S. Soatto, "Class segmentation and object localization with superpixel neighborhoods," in *Proc. Int'l Conf. Computer Vision*, 2009.
[65] M. Marszalek and C. Schmid, "Accurate object recognition with shape masks," *Int. J. Computer Vision*, 2012.

Kuang-Jui Hsu received the B.S. degree from the Department of Electrical Engineering, National Sun Yat-sen University, and the M.S. degree from the Graduate Institute of Networking and Multimedia, National Taiwan University, in 2011 and 2013, respectively. Currently, he's a Ph.D. candidate in the Department of Computer Science and Information Engineering, National Taiwan University. He also serves as a research assistant in the Research Center for Information Technology Innovation, Academia Sinica. His research interests include computer vision, machine learning, deep learning, and image processing.

Yen-Yu Lin (M'12) received the B.B.A. degree in information management, and the M.S. and Ph.D. degrees in computer science and information engineering from National Taiwan University, Taipei, Taiwan, in 2001, 2003, and 2010, respectively. He is currently an Associate Research Fellow with the Research Center for Information Technology Innovation, Academia Sinica, Taipei, Taiwan. His current research interests include computer vision, machine learning, and artificial intelligence.

Yung-Yu Chuang received his B.S. and M.S. from National Taiwan University in 1993 and 1995 respectively, and the Ph.D. from University of Washington at Seattle in 2004, all in Computer Science. He is currently a professor with the Department of Computer Science and Information Engineering, National Taiwan University. His research interests include computational photography, computer vision and rendering.