國立臺灣大學電機資訊學院資訊工程學系暨研究所

博士論文

Department of Computer Science and Information Engineering
College of Electrical Engineering and Computer Science
National Taiwan University

Doctoral Dissertation

弱監督式卷積神經網路學習於視覺顯著物體發掘

Visual Attention-getting Object Discovery via Learning Weakly Supervised CNNs

許洸睿

Kuang-Jui Hsu

指導教授：莊永裕博士和林彥宇博士

Advisor: Yung-Yu Chuang, Ph.D & Yen-Yu Lin, Ph.D

中華民國 108 年 7 月

July, 2019

國立臺灣大學博士學位論文
口試委員會審定書

弱監督式卷積神經網路學習於視覺顯著物體發掘

Visual Attention-getting Object Discovery via Learning Weakly Supervised CNNs

本論文係許洸睿君（學號 D03922003）在國立臺灣大學資訊工程學系完成之博士學位論文，於民國 108 年 7 月 2 日承下列考試委員審查通過及口試及格，特此證明

口試委員：莊永裕　林秀宇

（指導教授）

陳祝嵩　[illegible]

陳煥宗　王鈺強

孫民　陳駿丞

賴尚宏

系　主　任　莊永裕

誌謝

首先，由衷地感謝我的兩位指導教授，台大資工莊永裕老師跟中研院資創中心林彥宇老師。我能有如此成就，有如此的成績，能完成這本博士論文，這兩位指導教授功不可沒。學生許洸睿三鞠躬感謝兩位指導教授的幫忙。第二，謝謝所有口委(名單詳見口試委員審定書)，願意抽時間參加博士論文考試，也提供許多建議，可以使我的研究更加完善，最重要的是，欣賞我的博論，可以讓我高分通過考試。第三，我非常感謝博士班這五年所遇到的實驗室夥伴，如 Tsun-Yi Yang、Yang-Ming Yeh、Yi-Hsuan Huang、Ting-Kuei Hu 、Chung-Chi Tsai、Chu-Ya Yang、Jo-Han Hsu、Ya-Fang Shih、Hao-Hsuan Chang、An An Yu、Han-Yi Wang、Yi-Ting Chen、Yun-Chun Chen、Han-Yi Lin、Yu-Lun Liu、Yi-Wen Chen、Cheng-Chun Hsu 等 (名單太長，請詳見 http://cvlab.citi.sinica.edu.tw/people)，外加隔壁實驗室但時常串門聊天的 Julia Chang，由於跟他們打屁聊天跟研究切磋，我才能度過這無聊五年歲月並有好的研究成果。第四，感謝我的大哥跟大嫂，能在我博士生涯，經濟最困難的時候，能體諒我並提供我金援協助，在我需要幫助時，伸出豬蹄援助我，弟弟這輩子絕對不會忘記。第五，非常謝謝我的國中死黨們，在我生涯最低點，願意陪時間打屁聊八卦，排解憂鬱，也在我需要幫忙時幫忙我。第七，感謝我在博士生涯中，拜過的所有神明跟廟宇，能讓我安然渡過博士生涯中的每一天，也讓我所有投稿都能順利。第八，感謝我在博士生涯中，所有幫忙過我的人。第九，也是最重要的，我非常謝謝爸爸媽媽這些年的照顧跟支持，從我小時候就願意花時間花栽培我開導我，同時也不

在意我像個米蟲需要被飼養，讓我無後顧之憂可以完成博士學業，我如今有所成就，主要歸功於我的父母。

摘要

在影像處理跟電腦視覺領域中，視覺顯著物體發掘是一個重要的且待解決的難題。這一個議題的主要目標是產生機率圖，用來標示影像中所有吸引人的區域。因為這個任務所產生的結果可以用來指出物體位置跟過濾不相關的背景，因此這個任務對於其他應用，如圖像重定位，視覺追蹤，物體切割跟辨認，是非常重要的。由於卷積神經網路可以有效果地學習影像特徵跟非線性分類器，因此目前的最佳方法都是建構在卷積神經網路上。不過，最大的缺點在於需要大量人工標記的像素等級的訓練資料訓練卷積神經網路。不過由於蒐集這類的資料需要花費量的人力資源，因此會限縮此類任務在其他應用的可能性。

這個博士論文中，我們提出四個方法處理上述所遇到的問題。第一個方法中，我們把類別特定的資訊加入視覺顯著物體發掘任務中，並且提出一個弱監督學習的方法來減少所需的標記成本。所提出的方法包含兩個以卷積神經網路為基的模組，影像等級的分類器跟像素等級的產生器，外加上四個損失函數。結果證明所提出的方法超越目前的弱監督跟強監督方法。第二個方法中，我們提出一個非監督式端對端訓練的方法叫做共注視卷積神經網路用於物體協同切割，且共注視卷積神經網路可以有效地獲得不同影像間物體一致性資訊，因此所提出的方法可以有效地超越目前的監督跟非監督式最佳方法。第三個方法中，我們提出一個新的卷積神經網路可以一起維持多張影像間物體間的一致性跟單張影像內物體的顯著性於物體共顯著偵測。所提出的方法超越最佳非監督式方法，也跟最佳監督式方法達到相同的效果。最後一個方法中，我們提出一個新的但困難的任務叫做實例等級物體協

同切割，並針對這個問題，提出一個共尖峰的概念用於定位且切割不同影像間的物體實例。我們利用共尖峰的概念發展一個簡單且有效的卷積神經網路為基架構用於這一個新的任務。我們針對這個新任務蒐集四個資料集，且在這四個資料集上，我們所提出的架構達到最佳效果。

關鍵字： 顯著性物體偵測，物體共顯著偵測，物體協同切割，卷積神經網路，非監督式學習，弱監督式學習

Abstract

Visual attention-getting object discovery has been an active topic in the fields of image processing and computer vision for decades. In this topic, the goal is to produce the saliency maps which highlights the regions of objects attracting people. This task is crucial to various applications such as image retargeting, visual tracking, object segmentation, and object recognition because the produced results can indicate objects of interest and mask out the irrelevant background. The current state-of-the-art saliency detection methods adopt convolutional neural networks (CNNs) because it has demonstrated effectiveness in joint visual feature extraction and nonlinear classifier learning. However, they require additional training data in the form of pixel-wise annotations, often manually drawn or delineated by tools with intensive user interaction. Such heavy annotation cost makes these methods less practical as pointed out in other applications.

In this dissertation, we address the aforementioned issue by proposing four methods for visual attention-getting object discovery. In the first work, we integrate the class-specific information into the visual attention-getting object discovery and then propose a weakly supervised learning method to reduce the annotation cost. The proposed method is composed of two CNN-based modules, image-level classifier and a pixel-level map generator with four losses. The results show that our approach outperforms the state-of-the-art weakly supervised methods and many fully supervised ones in both accuracy and efficiency. In the second work, we address unsupervised CNN-

based object co-segmentation under an end-to-end trainable scheme and thus propose a co-attention CNNs to explore the inter-object consistency. The proposed method remarkably outperforms the state-of-the-art unsupervised and supervised methods on the standard object co-segmentation benchmarks. In the third work, we focus on unsupervised CNN-based object co-saliency detection and propose an end-to-end trainable graphical CNNs to jointly preserve the inter-object consistency and explore the intra-object saliency. The results show that our approach remarkably outperforms the state-of-the-art unsupervised methods and even surpasses many supervised DL-based saliency detection methods. In the final work, we tackle the CNN-based instance co-segmentation, which is a new and challenging task, and propose the concept, co-peak, to localize and segment each object instance in the given images. We develop a simple and effective method is developed for instance co-segmentation. The proposed method learns a model based on the *fully convolutional network* (FCN) by optimizing three proposed losses. The learned model can reliably detect co-peaks and co-saliency maps for instance mask segmentation. Four datasets are collected for evaluating instance co-segmentation, and we achieve the state-of-the-art performance on these four datasets.

Keywords: saliency object detection, object co-saliency detection, object co-segmentation, convolutional neural networks, unsupervised learning, weakly supervised learning

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Discovering the visual attention-getting or salient[1] objects has attracted significant attention in the field of computer vision. The goal of the salient object discovery is to generate the real-valued probability maps or binary segment masks which highlight the regions of objects attracting people. As an essential component of image analysis and scene understanding, they are crucial and necessary to various computer vision applications, such as visual tracking [50], object recognition [119] and image retargeting [172], since objects of interest are kept while the irrelevant background is filtered out.

Recently, the state-of-the-art methods for salient object discovery are usually built on the *convolutional neural networks* (CNNs) [85] because CNNs have demonstrated the effectiveness in joint visual feature extraction and nonlinear mapping learning. However, training a reliable CNN model requires a lot of training data in the form of pixel-wise annotations, which are usually manually drawn or delineated by tools with intensive user interaction. The massive annotation cost of training data collection hinders the advances in attention-getting object discovery. In this thesis, we tackle this problem by using different cues, image-level labels and mutual reference across different images, to jointly learn a CNN model and discover the salient objects. We address four tasks including top-down saliency detection, co-segmentation, and co-saliency detection. The first one considers only the image-level labels, and the latter three ones use the mutual reference across different images.

[1] In the following thesis, the term "salient" is used instead of the term "attention-getting."

1.1 Thesis overview

This thesis consists of four parts, and in each of them, we address one specific task related to the visual attention-getting object discovery. In the following subsections, each task is introduced, including the goals, difficulties, and contributions.

1.1.1 A Category-Driven Map Generator for Single-image Saliency Detection

The methods for visual attention-getting object discovery in a single image (or object saliency detection for short) can be roughly divided into the bottom-up and the top-down groups. The bottom-up methods [21, 69, 70, 138] 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 [23, 25, 48, 80, 155] are category-aware. They utilize the prior knowledge about a target object category for saliency detection and do not suffer from the limitation as mentioned above. However, they require the training data in the form of pixel-wise annotations to learn the prior knowledge, and the massive annotation cost of training data collection hinders the advances.

In this task, 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., [23, 24], our approach carries out top-down saliency detection based on *convolutional neural networks* (CNNs) [85]. 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

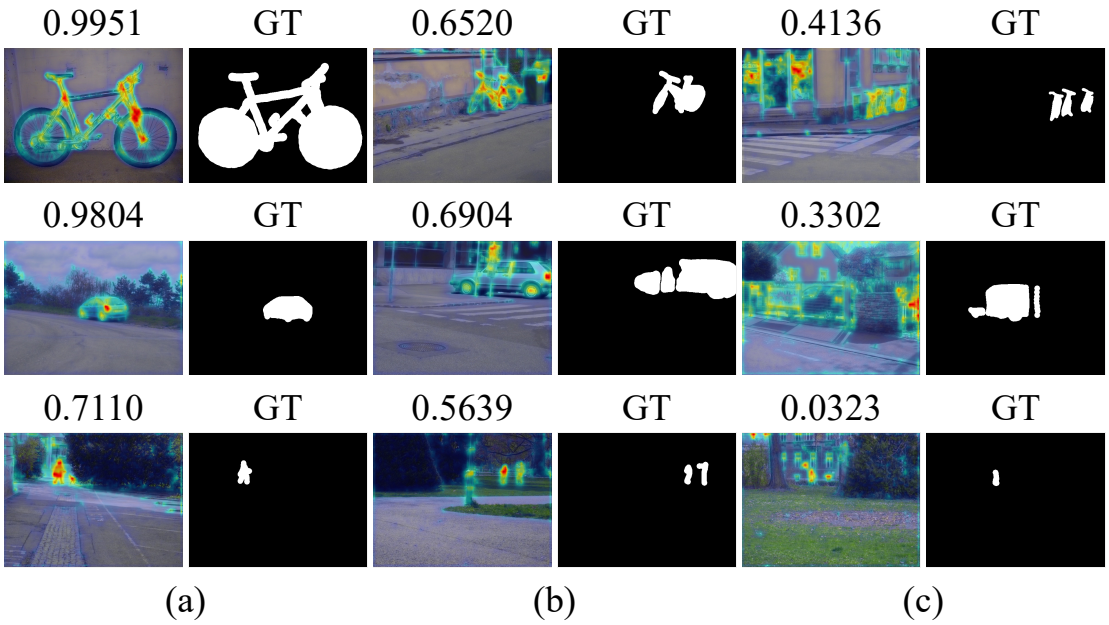


Figure 1.1: Examples of the detected saliency maps and their ground truth annotations for the categories *bike* (top row), *car* (middle row), and *person* (bottom row). On the top of each map, we show the score by applying the classifier to the image with its non-salient regions removed. For images 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.

the same time, 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 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. Figure 1.1 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.

Contributions. The main contribution of this work is to develop a 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 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 [113] and PASCAL VOC-07/12 [30]. 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.

1.1.2 Co-Attention CNNs for Object Co-Segmentation

Visual attention-getting object segmentation across multiple images (or object co-segmentation for short) simulates human visual systems to search for the common objects repetitively appearing in images. It was introduced in [122] to address the difficulties of single-image object segmentation. It leverages not only intra-image appearance but also inter-image object co-occurrence to compensate for the absence of supervisory information.

Engineered features, such as SIFT [107], HOG [28], and texton [129], are widely used in conventional co-segmentation methods, e.g., [74, 86, 135, 142], to cope with intra-class variations and background clutters. These features are designed in advance. The feature extraction and object co-segmentation are treated as separate steps leading to suboptimal performance because they are not optimized for the given images for co-segmentation. Yuan et al. [158] proposed a CNN-based supervised method, which learns the mapping between images and the corresponding masks, for object co-segmentation. They achieved the state-of-the-art results by substituting the features learned by CNNs for engineered features. However, their method requires additional training data in the form of object masks for learning the CNN model.

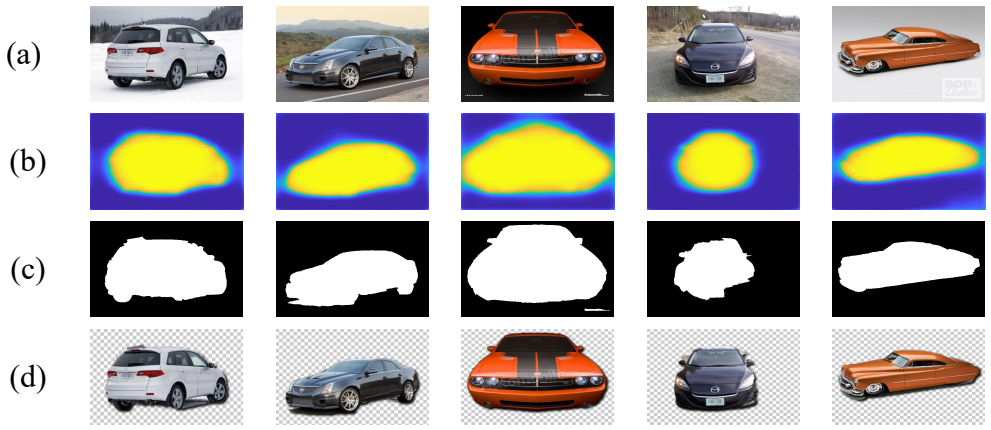


Figure 1.2: (a) The images for co-segmentation. (b) The estimated object maps by optimizing the co-attention loss. (c) The selected object proposals by using the mask loss. (d) Our co-segmentation results by considering the two losses simultaneously.

This work presents a CNN-based method for co-segmentation that makes a good compromise between the performance and data annotation cost. Specifically, we aim at co-segmenting images covering objects of a specific category without additional data annotations. Our method does not rely on training data in the form of object masks and can improve co-segmentation by using the features extracted by CNNs. To this end, we develop the two losses, *co-attention loss* and *mask loss*, to derive a CNN model by enhancing the similarity among the estimated objects across images while enforcing the figure-ground distinctness in each image. Figure 1.2 shows an example of the co-segmentation results inferred by our method.

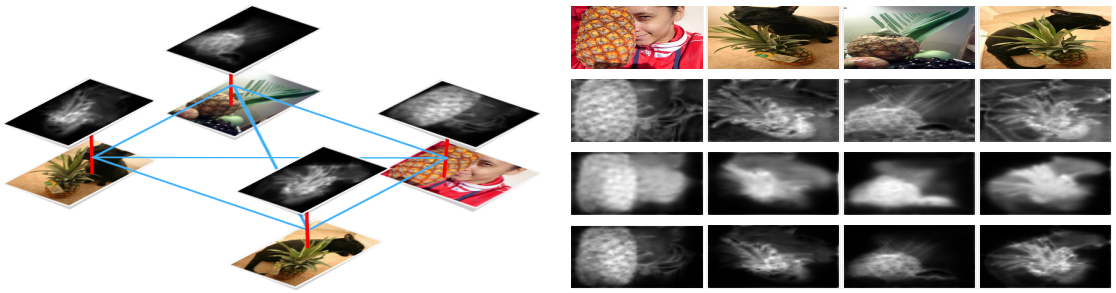
Contributions. This work is the first attempt to develop an end-to-end trainable CNN model for object co-segmentation without the pixel-wise annotations as training data. Compared with conventional unsupervised methods [74, 86, 135, 142] and the supervised CNN-based method [158], our approach can enjoy the boosted performance empowered by deep CNN features and does not suffer from the high annotation cost in labeling object masks as training data. Our method is evaluated on three benchmarks for co-segmentation, *the Internet dataset* [124], *the iCoseg dataset* [4], and *the PASCAL-VOC dataset* [31]. It remarkably outperforms the state-of-the-art unsupervised and supervised methods.

1.1.3 Graphical CNNs for Object Co-Saliency Detection

Visual attention-getting object saliency detection across multiple images (or object co-saliency detection for short) refers to searching for visually salient objects repetitively appearing in multiple given images. Different from the co-segmentation tasks in Section 1.1.2, the goal in this task is to generate a probability map which highlights the salient objects instead of a segmentation mask. Similar to the object co-segmentation, the success of conventional co-saliency detection also relies on robust feature representations to tackle the problem resulting from the large intra-class variability and subtle figure-ground discrimination, such as the fixed engineered features, DL-based features or the ones learning from the training dataset. Although the learning-based methods achieve the state-of-the-art performance, they require the pixel-wise annotations.

In this work, In this work, we propose a CNN-based method for joint adaptive feature learning and co-saliency detection for given images without pixel-wise annotations. In the proposed method, co-saliency detection is decomposed into two complementary parts, *single-image saliency detection* and *cross-image co-occurrence region discovery*. The former detects the saliency object in a single image, which may not repetitively appear across images. The latter discovers regions repetitively appearing across images, which may not be visually salient. To this end, we design two novel losses, *the single-image saliency (SIS) loss* and *the co-occurrence (COOC) loss*, to capture the two different but complementary sources of information. These two losses measure the quality of the saliency maps by referring to individual images and the co-occurrence regions for each image pair, respectively. They are further integrated on a graphical model whose unary and pairwise terms correspond to the proposed SIS and COOC losses respectively, as illustrated in Figure 1.3 (a). Through optimizing the proposed losses, our approach can generate co-saliency maps of high quality by integrating SIS and COOC cues, as shown in Figure 1.3 (b).

Contributions. To the best of our knowledge, our method is the first one that the CNN model for co-saliency detection is learned without the pixel-wise annotations as training



(a) (b)

Figure 1.3: Motivation of our method. (a) Our method optimizes an objective function defined on a graph where single-image saliency (SIS) detection (red edges) and across-image co-occurrence (COOC) discovery (blue edges) are considered jointly. (b) The first row displays the images for co-saliency detection. The following three rows show the detected saliency maps by using COOC, SIS, and both of them, respectively.

data. Compared with the conventional approaches including those using engineered features [35, 66, 105, 137, 138] and those using DL-based features [161, 163], our method achieves better performance by joint adaptive feature learning and co-saliency detection based on CNNs. Compared with the supervised methods [39, 160], our method can reach comparable or even slightly better performance and does not suffer from the high annotation cost of labeling object masks as training data. We comprehensively evaluate our method on three benchmarks for co-saliency detection, including *the MSRC dataset* [153], *the iCoseg dataset* [4], and *the Cosal2015 dataset* [161]. The results show that our approach remarkably outperforms the state-of-the-art unsupervised methods and even surpasses many supervised DL-based saliency detection methods.

1.1.4 CNNs for Instance-Level Object Co-Segmentation

As mentioned in Section 1.1.2, object co-segmentation has recently gained significant progress owing to the fast development of *convolutional neural networks* (CNNs). The CNN-based methods [57, 92, 158] learn the representation of common objects in an end-to-end manner and can produce object-level results of high quality. However, they do not explore instance-aware information, i.e., one segmentation mask for each instance rather than each class, which is more consistent with human perception and offers better image understanding, such as the locations and shapes of individual instances. Therefore, we

present a new and challenging task called *instance-aware object co-segmentation* (or *instance co-segmentation* for short), which is extension of the object co-segmentation. Two examples of this task are shown in Figure 1.4 for a quick start. Given a set of images of a specific object category with each image covering at least one instance of that category, instance co-segmentation aims to identify all of these instances and segment each of them out, namely one mask for each instance. Note that unlike semantic [54, 20, 12] or instance segmentation [171], no pixel-wise data annotations are collected for learning. The object category can be arbitrary and unknown, which means that no training images of that category are available in advance.

This task is important since instance-level segmentation is preferable for humans and many vision applications. It is also challenging because no pixel-wise annotated training data are available and the number of instances in each image is unknown. We solve this task by dividing it into two sub-tasks, *co-peak search* and *instance mask segmentation*. In the former sub-task, we develop a CNN-based network to detect the co-peaks as well as co-saliency maps for a pair of images. A co-peak has two endpoints, one in each image, that are local maxima in the response maps and similar to each other. Thereby, the two endpoints are potentially covered by a pair of instances of the same category. In the latter sub-task, we design a ranking function that takes the detected co-peaks and co-saliency maps as inputs and can select the object proposals to produce the final results. Our method for instance co-segmentation and its variant for object co-localization are evaluated on four datasets, and achieve favorable performance against the state-of-the-art methods.

Contributions. We make the following contributions in this work. First, we introduce a new and interesting task called instance co-segmentation. Its input is a set of images containing object instances of a specific category, and hence is easy to collect. Its output is instance-aware segments, which are desired in many vision applications. Thus, we believe instance co-segmentation worth exploring. Second, a simple and effective method is developed for instance co-segmentation. The proposed method learns a model based on the *fully convolutional network* (FCN) [106] by optimizing three proposed losses. The learned model can reliably detect co-peaks and co-saliency maps for

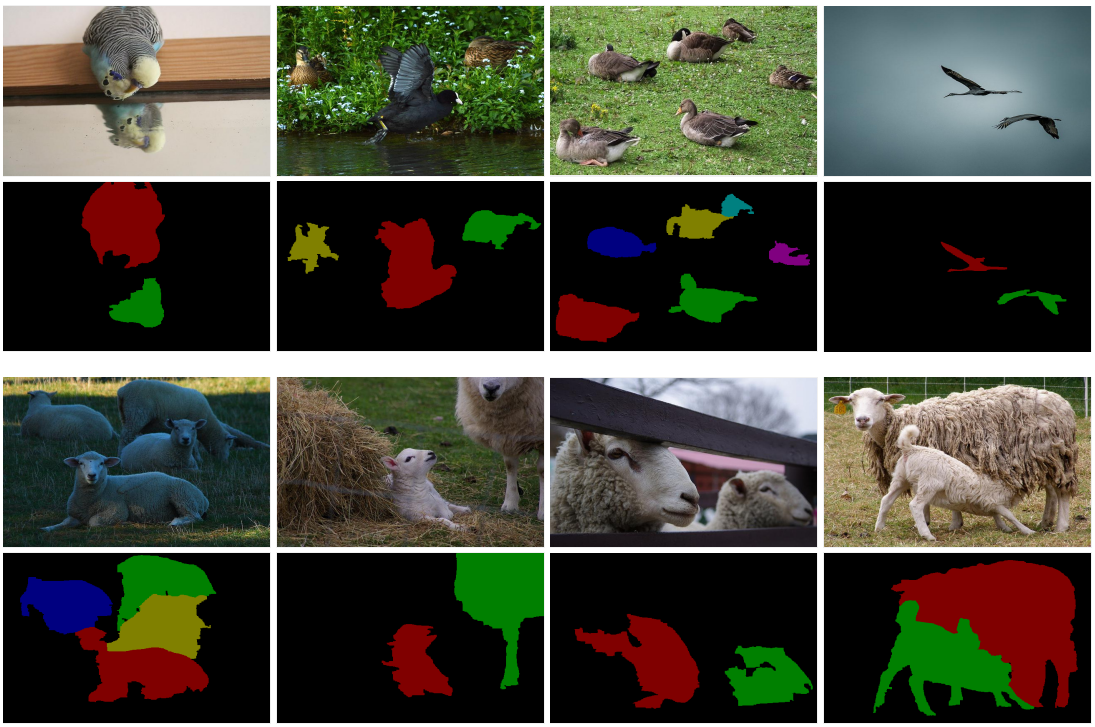


Figure 1.4: Two examples of instance co-segmentation on categories *bird* and *sheep*, respectively. An *instance* here refers to an object appearing in an image. In each example, the top row gives the input images while the bottom row shows the instances segmented by our method. The instance-specific coloring indicates that our method produces a segmentation mask for each instance.

instance mask segmentation. Third, we collect four datasets for evaluating instance co-segmentation. The proposed method for instance co-segmentation and its variant for object co-localization [22, 26, 134, 149, 150] are extensively evaluated on the four datasets. Our method performs favorably against the state-of-the-art methods.

1.2 Thesis organization

The remainder of the thesis is organized as follows. In Chapter 2, the weakly supervised top-down saliency detection is addressed, where only the image-level labels are utilized to learn the CNN model. In Chapter 3, we present an algorithm which integrates the mutual information across the different images into the CNNs, and thus the CNNs can be automatically optimized without any object masks. In Chapter 4, the graphical model built on the CNNs is adopted to capture the intra-image and inter-image information. In Chapter 5, we address a new and challenging task called *instance co-segmentation*, and

propose a simple CNN-based solution to solve it. Finally, in Chapter 6, we conclude and propose the potential extension.

Chapter 2

A Category-Driven Map Generator for Single-image Saliency Detection

In this chapter[1], we address the CNN-based weakly supervised top-down saliency detection. During the training stage, only the image-level labels are provided. As mentioned in Section 1.1.1, our proposed method is based on the observation: an image masked with a saliency map which highlights more object regions has higher the prediction scores from a pre-trained classifier. This observation is adopted to reduce the usage of pixel-wise annotated training data. Specifically, our approach is composed of two CNN-based modules, an *image-level classifier* and a *pixel-level map generator*, as shown in Figure 2.1. The classifier is trained with the image-level labels. 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.

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

[1]Published papers: [56, 59]

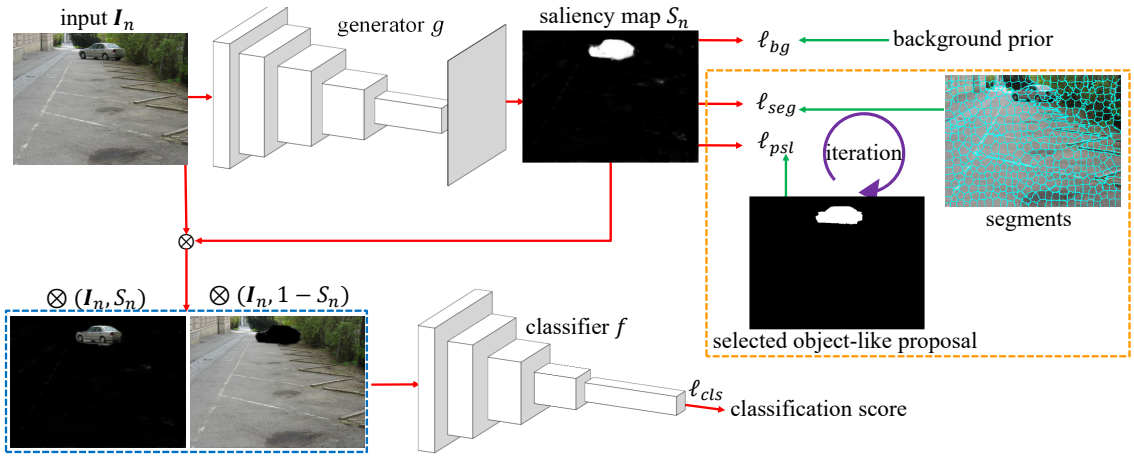


Figure 2.1: The overview of our approach. The classifier f distinguishes images of a target class from the rest. It propagates the classification information via the loss function ℓ_{cls} to train the generator g, which compiles saliency maps so that the masked images can be predicted by f with higher 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.

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.

In the experiments, we evaluate our approach on the three standard benchmarks for top-down saliency detection, Graz-02 [113], PASCAL VOC 2007 (VOC07) and PASCAL VOC 2012 (VOC12) [30]. It remarkably outperforms the state-of-the-art weakly supervised and many strongly supervised methods. In the following sections, we first give the literature review in Section 2.1. Next, we describe the problem definition and the proposed method in Section 2.2. Finally, the experiments are given in Section 2.3.

2.1 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.

2.1.1 Bottom-up object saliency detection

Bottom-up object saliency detection [7] receives much research attention owing to superior computational efficiency and less requirement of training data. As discussed in the survey paper [7], 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 [109, 100], global/local contrast [21, 69], focusness [70], objectness [14, 67, 70]. 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., [77, 143, 91], were proposed to capture the concept of objects, such as the space learning [77, 91] or a random forest regressor with contrast descriptors [143]. Recently, more and more researches [144, 146, 49, 93, 145, 167, 166, 102, 94] utilize CNNs to carry out the tasks of bottom-up object saliency detection in different ways, such as multi-level feature aggregation [166], uncertain convolutional feature learning [167], global context and local context integration [102], and contour-saliency conversion [94].

Wang et al. [145] 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.

2.1.2 Top-down object saliency detection

Top-down saliency methods such as [155, 80, 25, 23, 24] 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 [155] proposed a method for top-down saliency detection by jointly learning conditional random fields and a dictionary. Kocak et al. [80] computed the first and second order statistics and objectness on superpixels to distinguish target objects from the background. Cholakkal et al. [25] proposed the *locality-constrained contextual sparse coding* (LCCSC) method for top-down saliency detection. He et al. [48] 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. [23, 24] 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 [23] 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. [24] 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 [24] 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 [24] 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 [24] 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 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 [155, 80, 25, 23, 24], 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 [48], 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 [48] in the experiments.

This work shares the same network architecture with our prior work [56], 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 [56] and this work in Figure 2.2 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 [56].

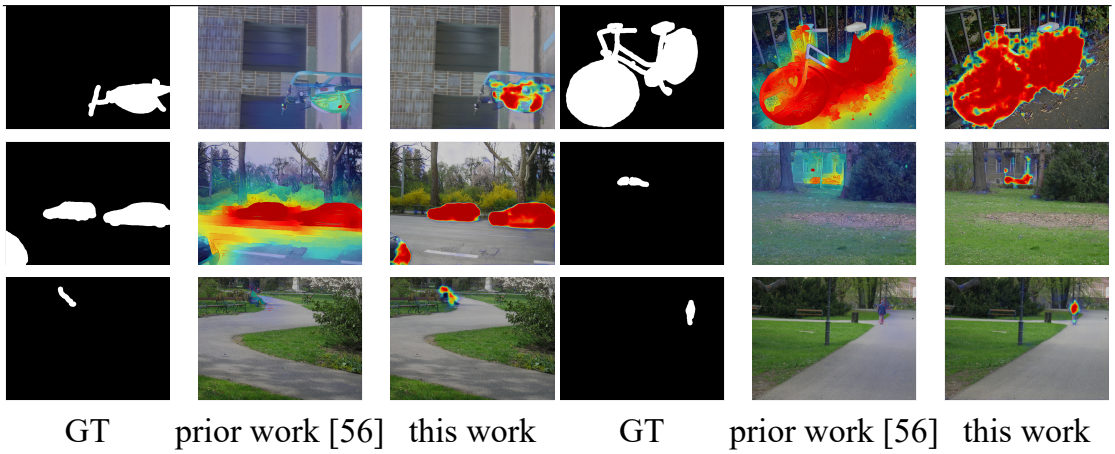


Figure 2.2: Comparison between the proposed method and our prior method. Each example consists of the ground truth (GT) and two saliency maps, generated by our prior work [56] and this work, respectively. Examples from three object categories, including *bike*, *car*, and *person*, are displayed in the three rows, respectively.

2.1.3 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 [87, 6, 75, 5, 169, 114]

and semantic segmentation [81, 152, 53, 128]. 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 [169, 114] are the most similar to our approach, because they generate saliency maps, too. The approach in [114] 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. [169] 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 [114] and [169] 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.

2.1.4 Top-down neural attention

Different from top-down object saliency detection, the methods in [130, 159, 9, 169, 164, 126] analyze the neuron responses or gradient of a classifier to generate class-specific activation maps. In [130, 159], the partial derivatives of neuron activations from error backward propagation are computed to highlight important image regions. In [9], 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. [169] 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. [164] 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) [106] 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.

2.2 Proposed approach

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

2.2.1 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 testing images.

2.2.2 Our formulation

As shown in Figure 2.1, 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

$$\begin{aligned}\ell(\mathbf{w}) = &\sum_{I_n \in D_{obj}} \ell_{cls}(I_n; \mathbf{w}) + \lambda_{seg}\ell_{seg}(I_n, M_n; \mathbf{w}) \\ &+ \lambda_{psl}\ell_{psl}(I_n, O_n; \mathbf{w}) + \sum_{I_n \in D_{bg}} \lambda_{bg}\ell_{bg}(I_n; \mathbf{w}),\end{aligned} \tag{2.1}$$

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.

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:

$$\begin{aligned}\ell_{cls}(I_n; \mathbf{w}) = &\|f(\otimes(S_n, I_n)) - 1\|^2 \\ &+ \|f(\otimes(1 - S_n, I_n)) - 0\|^2,\end{aligned} \tag{2.2}$$

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.

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

$$\ell_{bg}(I_n; \mathbf{w}) = \frac{1}{W \times H} \|S_n - Z\|^2, \tag{2.3}$$

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.

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:

$$\begin{aligned} \ell_{seg}(I_n, M_n; \mathbf{w}) = & \frac{1}{W \times H} \sum_{p \in M_n} \sum_{i \in p} \\ & \left[\frac{\sum_{j \in p} S_n(j)}{|p|} > 0.5 \right] \|S_n(i) - 1\|^2 \\ & + \left[\frac{\sum_{j \in p} S_n(j)}{|p|} \leq 0.5 \right] \|S_n(i) - 0\|^2, \end{aligned} \tag{2.4}$$

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. (2.4)), 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. (2.4)) can be expressed equivalently as the following matrix form:

$$\ell_{seg}(I_n, M_n; \mathbf{w}) = \frac{1}{W \times H} \|S_n - G_n\|^2, \tag{2.5}$$

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

$$G_n(i) = \begin{cases} 1, & \text{if } \frac{\sum_{i \in p} S_n(i)}{|p|} > 0.5, \\ 0, & \text{otherwise}, \end{cases} \tag{2.6}$$

where p is the superpixel containing the pixel i.

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.,

$$\ell_{psl}(I_n, O_n; \mathbf{w}) = \frac{1}{W \times H} \|S_n - O_n\|^2, \tag{2.7}$$

$$\text{where } O_n = \arg\min_{O \in \mathcal{O}_n} \|S_n - O\|^2. \tag{2.8}$$

In Eq. (2.8)), $\mathcal{O}_n$ 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. (2.7)). 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. (2.8)) to overcome issues of incomplete objects or objects with holes.

2.2.3 Optimization process

The objective in Eq. (2.1)) 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. (2.6)). The most consistent proposals $\{O_n\}$ are selected based on the generated saliency maps via Eq. (2.8)). 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. (2.1)). 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.

2.2.4 Implementation details

We implemented the proposed network based on MatConvNet [141]. ResNet-50 [48] is adopted as the image-level classifier $f(\cdot)$, because using other network architectures, such as AlexNet [85] or VGG-16/19 [131], sometimes results in the vanishing gradient problem. The two-class classifier $f(\cdot)$ is pre-trained on ImageNet [29] 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 [131] setting of FCN [106] 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. (2.1)) 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. (2.4)), the superpixel extraction algorithm SLIC [1] implemented in VLFeat [140] 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. (2.8)), we use the fast object proposal generation algorithm, *geodesic object proposal* (GOP) [84]. 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

Algorithm 1 The Optimization Procedure

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

Train the image classifier $f(\cdot)$; (Sec. 2.2.4)
Extract the superpixels for each image; (Sec. 2.2.4)
Compute the object proposals for each image; (Sec. 2.2.4)
Initialize the map generator $g(\cdot)$; (Sec. 2.2.4)
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. (2.6), $\forall I_n \in D_{obj}$;
    Update $\{O_n\}$ with $\{S_n\}$ via Eq. (2.8), $\forall I_n \in D_{obj}$;
    Optimize objective in Eq. (2.1) with $\{G_n\}$ and $\{O_n\}$;
    if convergence **then**
        Return $g(\cdot)$;
    end if
end for

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

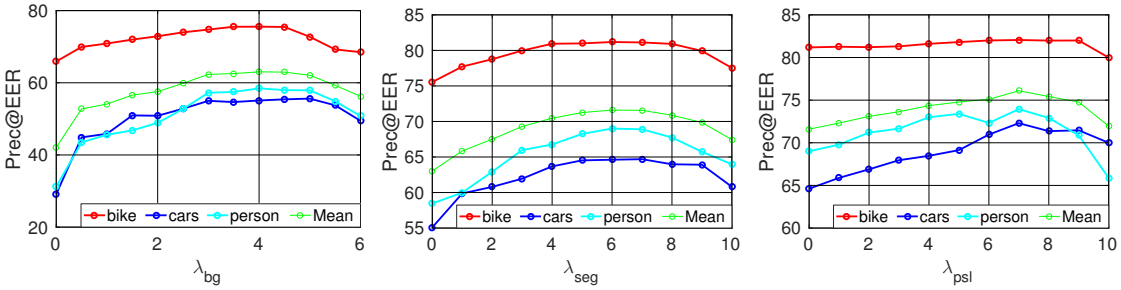
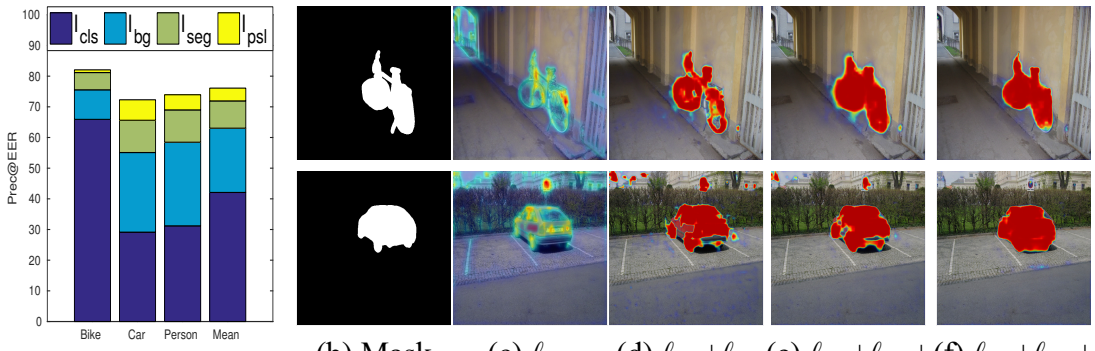


Figure 2.3: The performances of our approach in Prec@EER with different vales of weighting parameter (a) λ_{bg}, (b) λ_{seg}, and (c) λ_{psl} on the Graz-02 dataset, when adding the three loss functions ℓ_{bg}, ℓ_{seg}, and ℓ_{psl} one by one in the order.

parameters of SLIC and GOP are the same as those in their demo codes for superpixel extraction and unsupervised proposal generation, respectively.

2.3 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) Performance gain (b) Mask (c) ℓ_{cls} (d) $\ell_{cls}+\ell_{bg}$ (e) $\ell_{cls}+\ell_{bg}+\ell_{seg}$ (f) $\ell_{cls}+\ell_{bg}+\ell_{seg}+\ell_{psl}$

Figure 2.4: (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.

2.3.1 Datasets and evaluation criterion

We evaluated our proposed method on three benchmarks for top-down saliency detection, including Graz-02 [113], PASCAL VOC-07, and VOC-12 [30]. 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 [155, 80, 25, 23, 48, 56].

Graz-02. The Graz-02 dataset [113] 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 [155, 25, 23], 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.

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 [30] 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 [48, 114, 56], 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 [155, 25, 23], all models were evaluated on the 210 test images no matter whether the target objects are present or not.

Evaluation criterion. The *precision rate at equal error rate* (Prec@EER), was adopted to measure the performance. Following the previous researches [155, 80, 25, 23, 48, 56], 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. [48] to evaluate our model. In their work [48], 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.

2.3.2 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.

Model analysis. Following the previous work [155, 23], we analyze our model and empirically select the hyper-parameters on the Graz-02 test data. The proposed objective in Eq. (2.1)) 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},

Table 2.1: 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 [165]	US	54.7	39.0	52.0	48.6
	MST [139]	US	50.1	38.8	51.3	46.7
	WSS [145]	WS	64.7	71.6	64.0	66.8
	HDCT [77]	FS	55.9	43.8	53.0	50.9
	DRFI [143]	FS	51.3	49.6	59.6	53.5
	UCF [167]	FS	70.8	70.7	76.2	72.6
	Amulet [166]	FS	78.5	75.7	78.4	77.5
	PiCANet [102]	FS	79.7	82.1	85.0	82.3
	C2SNET [94]	FS	79.8	80.9	83.0	81.2
Top-down	ILC [2]	FS	71.9	64.9	58.6	65.1
	SP-Nei. [37]	FS	72.2	72.2	66.1	70.2
	Shape mask [110]	FS	61.8	53.8	44.1	53.2
	Patch-CRF [155]	FS	62.4	60.0	62.0	61.3
	SP-CRF [80]	FS	73.9	68.4	68.2	70.2
	LCCSC [25]	FS	76.2	71.2	64.1	70.5
	R-ScSPM [23]	FS	77.6	71.9	67.0	72.1
	R-ScSPM [23]	WS	67.5	56.5	57.6	60.5
	R-ScSPM+ [24]	WS	-	-	-	69.1
	Ours (prior)[56]	WS	78.9	66.6	64.2	69.9
	Ours	WS	82.1	78.5	75.0	78.5
	R-ScSPM+ [24]	WS+SP+SM	84.1	81.5	81.8	82.5

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. Figure 2.3(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 Figure 2.3(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 Figure 2.3(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 Figure 2.4(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 Figure 2.4(c) $\sim$ Figure 2.4(f). With only the classification loss ℓ_{cls}, the target objects, *bicycle* and *car*, are detected, but many false alarms occur in Figure 2.4(c). From Figure 2.4(c) to Figure 2.4(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 detected more confidently. From Figure 2.4(d) to Figure 2.4(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 Figure 2.4(e) to Figure 2.4(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 Figure 2.4(e), the detection result is incomplete since some holes are present inside the car. The unfavorable effect results from picking superpixels individually. In Figure 2.4(f) where the loss regarding object proposals has been incorporated, it is obvious that the object can be detected more completely.

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 2.1, where the field *setting* denotes the supervision condition of training data, including *unsupervised* (US), *weakly supervised* (WS), and *fully supervised* (FS) settings.

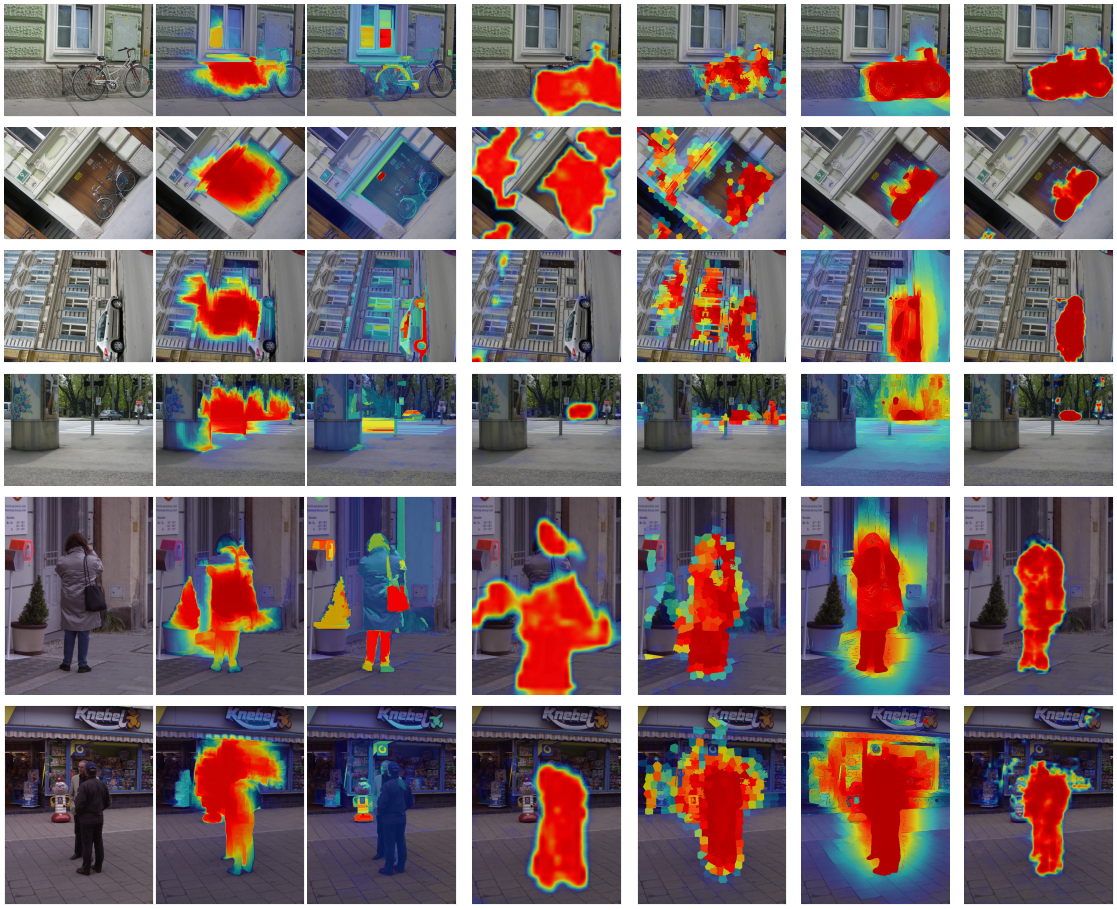


Figure 2.5: The saliency maps detected by our approach and the competing approaches on the Graz-02 dataset. In the six examples (rows), the target categories are *bike* in the first two rows, *car* in the middle two rows, and *person* in the last two rows.

Table 2.2: The average run time (in seconds) of the competing methods and our method on the Graz-02 dataset.

Method	MB [165]	MST [139]	Patch-CRF [155]	SP-CRF [80]
Time (Sec)	0.0263	0.1142	3.0940	30.2928
Speedup	1151.8×	265.3×	9.8×	1×

Method	Examplar[48]	Ours (prior) [56]	Exc. BP[164]	Ours
Time (Sec)	2.1470	5.2950	0.0632	0.0151
Speedup	14.1×	5.7×	479.3 ×	2006.1×

In Table 2.1, the conventional bottom-up methods [165, 139, 77, 143] 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 [167, 166, 102, 94] 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 [145] which also uses image-level labeled training data, our method reaches the much better performance. The main reason is that WSS [145] is a bottom-up method while our method handles top-down saliency detection and addresses salient objects of a target category.

Instead, fully supervised, top-down methods [2, 25, 37, 80, 110, 155] 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 [23] 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 [23] and our prior work [56] by large margins around 18% and 8.6% in Prec@EER, respectively. The large performance gain of our method over its prior work [56] 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+ [24] 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

Table 2.3: 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 [155]	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 [25]	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 [23]	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) [56]	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 [80]	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

where post-processing is turned off, our method outperforms R-ScSPM+ [24] 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, Figure 2.5 shows some detected saliency maps by different approaches. The bottom-up approaches, MB [165] and DRFI [143], 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 [155] and SP-CRF [80], 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 [155] or a superpixel [80] 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 [56], 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 Figure 2.5, our method does not suffer from the aforementioned issues. It can better preserve the object boundaries than the prior work [56] 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 out-

Table 2.4: 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 [145]	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 [166]	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 [167]	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 [102]	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 [94]	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+ [24]	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 [155]	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 [80]	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 [48]	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 [114]	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 [164]	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+ [24]	WS	50.2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Ours (prior) [56]	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

performs 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 [165, 139], non CNN-based top-down methods [155, 80], CNN-based top-down methods [48, 56], and top-down neural attention [164]. Note that extracting superpixels and object proposals is not required for our method during inference. Table 2.2 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. [48] 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 2.1. Nevertheless, we compare our method with their method in terms of running time, as shown in Table 2.2. 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 [106].

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 [155, 80] 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 [48] requires multiple forward passes and extra computation for the gradients via back-propagation [164]. Third, our method does not need any optimization or post-

processing process in the test stage. On the contrary, our prior work [56] 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 [165, 139].

2.3.3 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 2.3 and Table 2.4, 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 [145, 167, 166, 102, 94], two top-down saliency detection methods [155, 80], the state-of-the-art method [48], a method based on object localization [114], a method based on neural attention [164], and our prior work [56]. The competing methods [145, 48, 114, 164, 56] are also based on CNNs. The competing methods [155, 80, 48, 167, 166, 102, 94] adopt the fully supervised (FS) setting, while the others [145, 114, 164, 56] adopt the weakly supervised (WS) one.

In Table 2.4, our proposed method performs favorably against all competing methods. The WS localization methods [114, 164] 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 [114, 164] are significant, around 8.7% and 11.5% respectively. The WS bottom-up saliency method [145] 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+ [24] achieves better performance than our method. Nevertheless, under the setting where no post-processing steps are used, our method can outperform R-ScSPM+ [24]. Although adopting the weakly supervised setting, our method even achieves a slightly better performance than the state-of-the-art FS method [48]. 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 [167, 166, 102, 94], 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 [155, 80, 25] and the WS methods [23, 56]. The results are shown in Table 2.3. Note that our method is evaluated with two different experimental settings, one for the comparison with the method in [80] and the other for other methods: * in field *setting* of Table 2.3 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 2.3 and Table 2.4, our method provides significant gains over our prior work [56], 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.

Figure 2.6 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 [48] and the WS top-down method [164]. 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

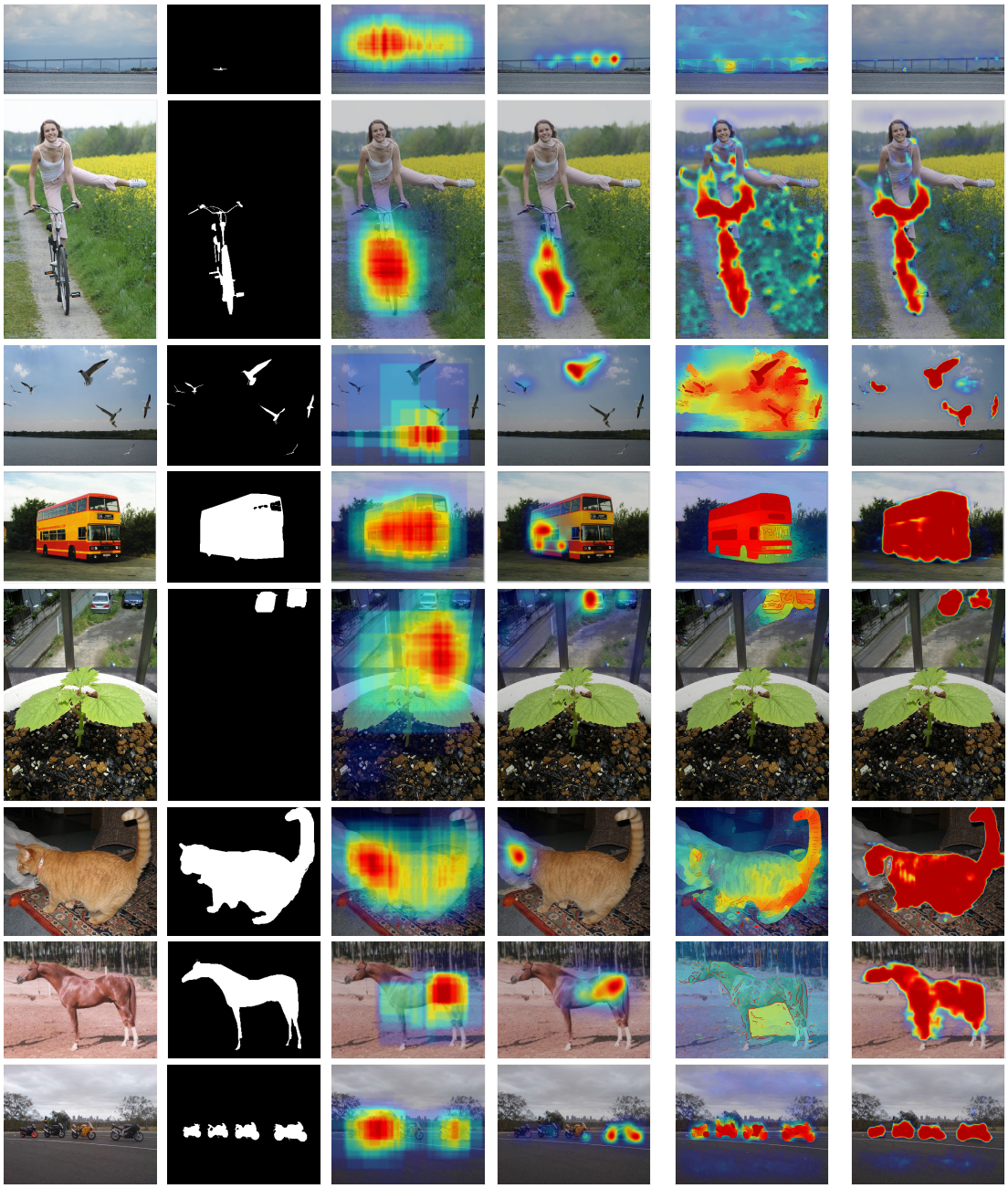


Figure 2.6: The saliency maps detected by different approaches on the PASCAL VOC-12 dataset. For top to bottom, the target object categories are *airplane, bicycle, bird, bus, car, cat, horse,* and *motorbike*, respectively.

Figure 2.7: 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.

(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, fifth and eighth rows. Although the FS method [48] achieves the performance comparable to ours, it tends to produce coarse saliency maps and can't preserve object boundaries. Compared to our prior work [56], 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.

2.3.4 Failure cases

We show some failure cases of our approach in Figure 2.7. 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.

Chapter 3

Co-Attention CNNs for Object Co-Segmentation

In this chapter[1], we address a CNN-based co-segmentation without pixel-wise annotation training, and hence the proposed method can make a good compromise between the performance and data annotation cost. More specifically, we focus on co-segmenting images consisting of objects of a specific category. Our method does not rely on training data in the form of object masks but enjoys the advantages of CNNs to boost the co-segmentation performance via the features optimized by CNNs. To this end, we develop the *co-attention loss* to derive a CNN model by enhancing the similarity among the estimated objects across images while enforcing the figure-ground distinctness in each image. Our model comprises two CNN modules, i.e., *a co-attention map generator* and *a feature extractor*, as shown in Figure 3.1. The generator compiles a heat map for the object in each image to estimate its figure-ground segmentation. The extractor computes the features of the estimated objects and backgrounds to minimize the co-attention loss. Through back-propagation, the generator is learned to compile high-quality object maps with which the resultant figure-ground segmentation can best optimize the co-attention loss. In this way, our model is end-to-end trainable and can carry out unsupervised object co-segmentation. For further enhancement, we develop the *mask loss*, which can refine the yielded object maps by preserving the whole objects and removing the noises. We evaluate the proposed

[1]Published papers: [57]

method on three benchmarks for co-segmentation, *the Internet dataset* [124], *the iCoseg dataset* [4], and *the PASCAL-VOC dataset* [31]. It significantly outperforms the state-of-the-art unsupervised and supervised methods.

In the following sections, we first give the literature review in Section 3.1. Next, we describe the proposed method in Section 3.2. Finally, the experimentS are presented in Section 3.3.

3.1 Related work

The literature related to our work is discussed in this section.

3.1.1 Object co-segmentation

According to [135], conventional researches on object co-segmentation can be divided into two categories, namely the *graph-based* [15, 125, 13, 66, 116, 142, 90] and the *clustering-based* [73, 76, 74, 86, 135] methods. The former methods adopt a structure model to capture the relationship between instances from different images and utilize the information shared cross images to select the most similar instances as the common objects jointly . The latter methods assume that the pixels or superpixels in the common objects can be grouped well. Thus, they formulate co-segmentation as a clustering problem to search for the common objects. In these graph-based and clustering-based methods, engineered features, e.g., SIFT, HOG, and texton, are often used for instance representation. The features are pre-designed instead of optimized for the input images. In contrast, our method adaptively learns the CNN features conditional on the given images. It can better cope with the intra-class variations and background clutters, leading to higher performance.

To improve the performance of co-segmentation, Sun and Ponce [132] further explored additional background images to help detect discriminative object parts. Yuan et al. [158] recently proposed a method integrating *conditional random fields* (CRFs) into CNNs to jointly learn the features and search the common objects. Despite the high performance, their method intensely relies on a large number of training object masks. It reduces the

applicability of their method to unseen images. Instead, the proposed method does not require additional background images or any training data but merely a set of images for co-segmentation. It can adapt itself to any unseen images in an unsupervised manner. Therefore, the proposed method has better generalization than the supervised method [158], and even outperforms it based on the developed co-attention and mask losses.

3.1.2 Unsupervised CNN for image correspondence

CNNs have been applied in an unsupervised fashion to a few tasks related to image correspondence, such as optical flow [157, 118, 111] and stereo matching [38, 170]. The common goal of these tasks is to find the cross-image correspondences of all pixels. The input images are typically adjacent video frames or stereo pairs of the same scene. The adopted objective functions are often based on brightness and cycle consistency. Namely, all matched pixels need to have similar colors or appearances, and the correspondences generated from different image perspectives should be consistent. There are three significant differences between these tasks and object co-segmentation. First, co-segmentation often considers objects of the same category, instead of the same instance. Thus, the brightness consistency may not hold. Second, co-segmentation identifies the region correspondence of the common objects, instead of the pixel correspondence of the whole image. Third, cross-image large displacement of the common objects may be present in co-segmentation. Local search for correspondence detection is no longer applicable. Due to the significant differences, these CNN-based methods for unsupervised image correspondence cannot be straightforwardly applied to object co-segmentation.

3.1.3 Weakly supervised semantic segmentation

Weakly supervised semantic segmentation (WSS) [81, 16, 72, 128, 52, 151, 123] aims to reduce the annotation cost of semantic segmentation. Methods of this category usually train their models by using training data with image-level labels, instead of pixel-level masks. There are at least two significant differences between WSS and co-segmentation. First, WSS typically consists of the training and testing phases. It requires weakly an-

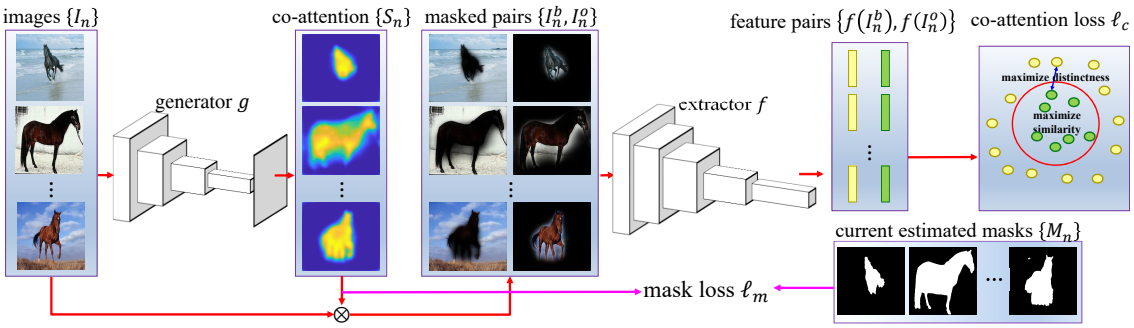


Figure 3.1: The overview of our method. Our network architecture is composed of two collaborative CNN modules, a map generator g and a feature extractor f, which are derived by the co-attention loss ℓ_c and the mask loss ℓ_m.

notated training data to learn the model and applies the learned model to test images. Co-segmentation is carried out by discovering objects commonly appearing in multiple images in a single phase. Second, WSS works with training images of various known categories and requires that the categories of testing images are covered by those of training images. On the contrary, co-segmentation usually works on multiple images of a single, unknown category.

3.2 Proposed approach

Our method is introduced in this section. First, the proposed formulation for co-segmentation is given. Then, the developed loss functions and the optimization process are described. Finally, some implementation details are provided.

3.2.1 Proposed formulation

Given a set of N images, $\{I_n\}_{n=1}^N$, commonly covering objects of the same category, our goal is to segment the common objects. Figure 3.1 illustrates the proposed method for a quick overview. Our network architecture is composed of two collaborative CNN modules, i.e., the co-attention map generator g and the semantic feature extractor f. Two loss functions, including the co-attention loss ℓ_c and the mask loss ℓ_m, are developed to derive the network.

The generator g is a *fully convolutional network* (FCN) [106]. For each image I_n, the generator estimates its co-attention map, $S_n = g(I_n)$, which highlights the common object in I_n. With S_n, the estimated object image I_n^o and background image I_n^b of I_n are available. The extractor f can be one of the pre-trained CNN models for image classification, such as AlexNet [85] or VGG-16 [131], with the softmax layer removed. It computes the semantic features of the estimated object and background images, i.e., $f(I_n^o)$ and $f(I_n^b)$. We treat the inputs to the last fully connected layer of f as the extracted features.

The co-attention loss ℓ_c is introduced to enhance both inter-image object similarity and intra-image figure-ground distinctness. The mask loss refines the co-attention maps by referring to the selected object proposals. It makes the maps retain the whole objects while removes the noises. According to our empirical studies, we pre-train the extractor f and fix it during training, although fine-tuning is possible. Suppose the generator g is parametrized by $\mathbf{w}$. The proposed unsupervised loss function for learning g is defined by

$$\ell(\mathbf{w}) = \ell_c(\{I_n\}_{n=1}^N; \mathbf{w}) + \lambda \sum_{n \in \{1,\dots,N\}} \ell_m(I_n, M_n; \mathbf{w}), \tag{3.1}$$

where λ is a constant for weighting losses. M_n is the selected object proposal for I_n. For the sake of clearness, the optimization of Eq. (3.1), the loss ℓ_c and the loss ℓ_m will be detailed in the following subsections.

From co-attention to co-segmentation. By applying the learned generator g to all images, the corresponding co-attention maps are obtained. Following [158], we generate the co-segmentation results via *dense CRFs* [83] where the unary and the pairwise terms are set to referring to the co-attention maps and bilateral filtering, respectively.

3.2.2 Co-attention loss ℓ_c

The co-attention loss ℓ_c guides the training of the generator g by referring to the object and background features computed by extractor f. This loss is designed based on the two criteria used in unsupervised object co-segmentation, namely high inter-image object similarity and high intra-image figure-ground distinctness.

As shown in Figure 3.1, the generator g produces the co-attention map S_n for each image I_n. Sigmoid function serves as the activation function in the last layer of g. Hence, the co-attention value at every pixel k, $S_n(k)$, ranges between 0 and 1. With S_n, the masked object and background images of I_n are respectively obtained as follows:

$$I_n^o = \otimes(S_n, I_n) \text{ and } I_n^b = \otimes(1 - S_n, I_n), \tag{3.2}$$

where $\otimes$ is the operator of element-wise multiplication. Images I_n^o and I_n^b highlight the estimated object and background of I_n, respectively.

The extractor f is applied to images $\{I_n^o, I_n^b\}_{n=1}^N$ for computing the features $\{f(I_n^o), f(I_n^b)\}_{n=1}^N$. With these features, the co-attention loss is then defined by

$$\ell_c(\{I_n\}_{n=1}^N; \mathbf{w}) = -\sum_{i=1}^{N}\sum_{j\neq i} \log(p_{ij}), \tag{3.3}$$

where p_{ij} can be considered as a score estimating two mentioned criteria of object co-segmentation, and it is defined by the following equations,

$$p_{ij} = \frac{\exp(-d_{ij}^+)}{\exp(-d_{ij}^+) + \exp(-d_{ij}^-)}, \tag{3.4}$$

$$d_{ij}^+ = \frac{1}{c}\|f(I_i^o) - f(I_j^o)\|^2, \text{ and} \tag{3.5}$$

$$d_{ij}^- = \frac{1}{2c}(\|f(I_i^o) - f(I_i^b)\|^2 + \|f(I_j^o) - f(I_j^b)\|^2). \tag{3.6}$$

Eq. (3.5) and Eq. (3.6) respectively measure the inter-image object distance and intra-image figure-ground discrepancy for an image pair I_i and I_j. Constant c is the dimension of the extracted features. The co-attention loss in Eq. (3.3) is defined over all image pairs. By minimizing this loss, the generator g will produce the co-attention maps in which low inter-image object distances and high intra-image figure-ground discrepancies can be observed. The co-attention loss is the primary part of the objective function. To the best of our knowledge, it has not been explored and is novel in the literature.

3.2.3 Mask loss ℓ_m

Using the co-attention loss alone may lead to two problems. First, the co-attention maps tend to highlight only the discriminative object parts, instead of the whole objects. It is not surprising since segmenting only the discriminative parts gives an even lower co-attention loss. Second, some noises, false positives here, are present in the co-attention maps.

The two problems can be alleviated by taking into account single-image objectness. To this end, we can compile a pool of object proposals, $\mathcal{O}_n$, for each image I_n, by using an unsupervised, off-the-shelf approach, e.g., [84]. These proposals are designed to cover objects completely. We can pick object proposals highly consistent with co-attention maps, and use them in order to regularize co-segmentation. Unfortunately, the co-attention maps $\{S_n\}$ at the early training stage are too unstable to pick satisfactory proposals. Thus, we adopt a two-stage strategy to optimize Eq. (3.1). At the first stage, the mask loss is turned off. After a few epochs, the resultant co-attention maps $\{\tilde{S}_n\}$ become stable enough to pick the proposals $\{\tilde{M}_n\}$, where $\tilde{M}_n = \arg\min_{O\in\mathcal{O}_n} \|\tilde{S}_n - O\|^2$. At the second stage, the mask loss ℓ_m in Eq. (3.1) is turned on and it is defined by

$$\begin{aligned}\ell_m(I_n, M_n; \mathbf{w}) = \frac{-1}{|\mathcal{K}|}\sum_{k\in\mathcal{K}}(\beta\tilde{M}_n(k) + (1-\beta)M_n(k))\log(S_n(k)) \\ + (\beta(1-\tilde{M}_n(k)) + (1-\beta)(1-M_n(k)))\log(1-S_n(k)),\end{aligned} \tag{3.7}$$

where $M_n = \text{argmin}_{O\in\mathcal{O}_n} \|S_n - O\|^2$, β is a constant, $\mathcal{K}$ is the index set of pixels, and $|\mathcal{K}|$ is the number of pixels. The cross entropy loss is adopted in Eq. (3.7), and enforces the co-attention map S_n to be consistent with the weighted combination of $\tilde{M}_n$ and M_n.

The idea behind the mask loss is intuitive: The object proposal, covering the discriminative parts detected by S_n, likely includes the non-discriminative parts at the same time. This property is leveraged to enforce the generator g to highlight the non-discriminative parts along with the discovered discriminant parts. The loss also reduces false positives because it can suppress the unfavorable high co-attention values in the background. The mask loss is inspired by the bootstrapping loss in [117], but with the difference that the estimated co-attention maps $\{S_n\}$ are updated in turn with the selected proposals instead of a hard threshold 0.5. β is set as 0.95 following [117].

Object mask refinement. An object proposal is designed to cover one single object. For an image where multiple objects are present, the mask loss as mentioned above may lead to an unfavorable circumstance. Namely, only one single object is detected. Thus, we develop a scheme to generate an object mask M_n by an iterative refinement procedure where multiple object proposals may be iteratively merged into M_n. Let O_n^t denote the selected proposal for image I_n at the tth iteration. At the first iteration, we pick the proposal O_n^1 from $\mathcal{O}_n$ that best matches the co-attention map S_n. The object mask M_n is initially set to O_n^1. Other proposals overlapping O_n^1 are removed from $\mathcal{O}_n$. The co-attention values in S_n are set to zero if the values are less than the average value of O_n^1. At the following iteration t, we pick proposal O_n^t that best matches the updated S_n, and merge it into M_n. Then the proposal pool $\mathcal{O}_n$ and the co-attention map S_n are similarly updated. The procedure is repeated until S_n becomes a zero matrix or no proposals remain in $\mathcal{O}_n$. This iterative scheme allows the object mask M_n to cover multiple non-overlapping and high-quality object proposals. The updated M_n is then substituted for the original M_n in Eq. (3.7).

3.2.4 Optimization process

The objective function in Eq. (3.1) is differentiable and convex. We choose Adam [78] as the optimization solver for its rapid convergence. In each epoch, we perform forward propagation and get the updated co-attention maps $\{S_n\}$. Then, the most consistent object masks $\{M_n\}$ are generated based on the proposed object mask generation scheme. Once the object masks $\{M_n\}_{n=1}^N$ are determined, the objective function in Eq. (3.1) can be optimized by using Adam. The gradients of each loss function with respect to the optimization variables can be derived straightforward. Therefore, we omit their derivation here.

Our method is end-to-end trainable. Feature extractor can be updated via back propagation. We keep it fixed because, for co-segmentation, there are often not sufficient images for the stable update. Besides, object proposals are dynamically refined to cover common objects better.

3.2.5 Implementation details

The proposed method is implemented based on $\mathrm{MatConvNet}$ [141]. The same network architecture is used in all the experiments. ResNet-50 [47] is adopted as the feature extractor f because AlexNet [85] and VGG-16/19 [131] sometimes lead to the problem of gradient vanishing in our cases. The feature extractor f is the off-the-shelf model pre-trained on ImageNet [29]. It is fixed during the optimization process. We have tried to fine-tune f based on the co-attention loss. The performance is not improved due to the limited number of images for co-segmentation. Thus, the feature extractor f remains fixed in the experiments. The features extracted by f are set to the inputs to the last fully connected layer of f. The feature dimension, i.e., c in Eq. (3.5) and Eq. (3.6), is set to $2{,}048$.

The generator g is developed based on the VGG-16 [131] setting of FCN [106]. We replace the activation function *softmax* in the last layer with the *sigmoid* function. The output of the sigmoid function serves as the co-attention map. The learning rate is set to 10^{-6} and kept fixed during optimization. As mentioned previously, the generator is learned in a two-stage manner. At the first stage, we optimize the objective in Eq. (3.1) with the mask loss ℓ_m removed for 20 epoches. After the first stage, the co-attention maps $\{\tilde{S}_n\}$ become stable enough to pick plausible $\{\tilde{M}_n\}$. At the second stage, the mask loss ℓ_m is turned on and the objective in Eq. (3.1) is optimized for 40 epochs. Therefore, the total number of epoches is 60. The batch size, weight decay, and momentum are set to 5, 0.0005, and 0.9, respectively. All images for co-segmentation are resized to the resolution 384×384 in advance since the feature extractor f is applied to only images of the same size. Then, we resize the generated co-segmentation results into their original sizes for the performance measure. The parameter λ in Eq. (3.1) is empirically set and fixed to 9 in all experiments.

For generating the pool of object proposals $\{\mathcal{O}_n\}$ used for object mask update, we adopt the fast object proposal generation algorithm, *geodesic object proposal* (GOP) [84]. Following the unsupervised setting in this work, the unsupervised setting of GOP is adopted. The number of the generated proposals for an image typically ranges from 200 to $1{,}100$.

Table 3.1: The performance of object co-segmentation on the Internet dataset. The numbers in red and green respectively indicate the best and the second best results. * means the supervised method.

Method	Airplane		Car		Horse		Avg.	
	$\mathcal{P}$	$\mathcal{J}$	$\mathcal{P}$	$\mathcal{J}$	$\mathcal{P}$	$\mathcal{J}$	$\mathcal{P}$	$\mathcal{J}$
[74]	47.5	0.12	59.2	0.35	64.2	0.30	56.97	0.243
[124]	88.0	0.56	85.4	0.64	82.8	0.52	82.73	0.427
[18]	90.2	0.40	87.6	0.65	89.3	0.58	89.03	0.543
[13]	72.6	0.27	75.9	0.36	79.7	0.36	76.07	0.330
[86]	52.8	0.36	64.7	0.42	70.1	0.39	62.53	0.392
[66]	90.5	0.61	88.0	0.71	88.3	0.61	88.93	0.643
[116]	91.0	0.56	88.5	0.67	89.3	0.58	89.60	0.603
[43]	77.7	0.33	62.1	0.43	73.8	0.20	71.20	0.320
[135]	79.8	0.43	84.8	0.66	85.7	0.55	83.43	0.547
[132]	88.6	0.36	87.0	0.73	87.6	0.55	87.73	0.547
[65]	81.8	0.48	84.7	0.69	81.3	0.50	82.60	0.556
w/o ℓ_m	93.6	0.66	91.4	0.79	87.6	0.59	90.86	0.678
Ours	94.2	0.67	93.0	0.82	89.7	0.61	92.29	0.698
[158]*	92.6	0.66	90.4	0.72	90.2	0.65	91.07	0.677

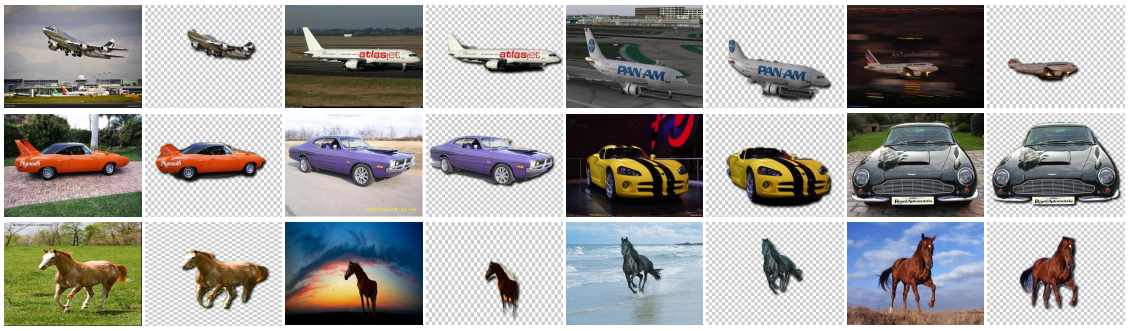


Figure 3.2: The co-segmentation results generated by our approach on the Internet dataset. In the three examples (rows), the common object categories are airplane, car, and horse, respectively.

3.3 Experimental results

In this section, we evaluate the proposed method and compare it with existing methods on three benchmarks for object co-segmentation, including the Internet dataset [124], the iCoseg dataset [4], and the PASCAL-VOC dataset [31]. These datasets are composed of real-world images with large intra-class variations, occlusions, and background clutter. They have been widely adopted to evaluate many existing methods for object co-segmentation, such as [66, 142, 158].

3.3.1 Datasets and evaluation metrics

The Internet dataset. This dataset introduced in [124] contains images of three object categories including airplane, car, and horse. Thousands of images in this dataset were collected from the Internet. Following the same setting of the previous work [124, 158, 135], we use the same subset of the Internet dataset where 100 images per class are available.

The iCoseg dataset. There are 38 categories in the iCoseg dataset [4] with total 643 images. Each category consists of several images, and these images contain either the same or different object instances of that category. Significant variations of viewpoints and deformations are present in this dataset.

The PASCAL-VOC dataset. This dataset was collected by Faktor and Irani [31]. It contains total $1{,}037$ images of 20 object classes from PASCAL-VOC 2010 dataset. The PASCAL-VOC dataset is more challenging and difficult than the Internet dataset due to extremely large intra-class variability and subtle figure-ground discrimination. Besides, some object categories have only a few images.

Evaluation metrics. Two widely used measures, *precision* ($\mathcal{P}$) and *Jaccard index* ($\mathcal{J}$), are adopted to evaluate the performance of object co-segmentation. Precision measures the percentage of correctly segmented pixels including both object and background pixels. Jaccard index is the ratio of the intersection area of the detected objects and the ground truth to their union area. The background pixels are taken into account in precision, so the images with larger background areas tend to have a higher performance in precision. Therefore, precision may not very faithfully reflect the quality of object co-segmentation results. Compared with precision, Jaccard index is considered more reliable to measure the quality of results. It provides the more appropriate evaluation as it only focuses on objects.

Table 3.2: The performance of object co-segmentation on the iCoseg dataset. The numbers in red and green respectively indicate the best and the second best results. * means the supervised method.

Method	[66]	[116]	[135]	[142]	Ours	[158]*
$\mathcal{P}$	91.8	93.3	90.8	93.8	96.5	94.4
$\mathcal{J}$	0.72	0.76	0.74	0.77	0.84	0.82

3.3.2 Comparison with co-segmentation methods

We compare the proposed method with the state-of-the-art methods on the Internet, iCoseg, and PASCAL-VOC datasets, and report their performances in Table 3.1, Table 3.2, and Table 3.3, respectively. All methods in Table 3.1, Table 3.2, and Table 3.3 are unsupervised except for the method [158]. Our approach achieves the state-of-the-art performance on the three datasets under both evaluation metrics.

On the Internet dataset, our method outperforms the unsupervised state-of-the-art [66] by a margin of around 5% in $\mathcal{J}$ and the supervised method [158] by a margin of around 2% in Table 3.1. On the iCoseg dataset, our method performs favorably against the state-of-the-art unsupervised method [142] by a margin of around 7% in $\mathcal{J}$ and the supervised method [158] by a margin of around 2% in Table 3.2. The results demonstrate that the proposed method can effectively utilize the information shared between common objects in different images without using complex graphical structures and optimization algorithms, or additional training data in the form of object masks. The effectiveness of our method mainly results from two properties. First, the co-attention loss enables CNNs to adaptively learn the robust features for unseen images, and discover the common regions. Second, the mask loss helps CNNs discover the whole objects and remove noises. Figure 3.2 and Figure 3.3 show some co-segmentation results on the Internet and iCoseg datasets, respectively. Our method can generate promising object segments under different types of intra-class variations, such as colors, sharps, views, and background clutter, in the Internet and iCoseg datasets, respectively.

In Table 3.3, our proposed method also outperforms the best competing method [142] by a large margin of around 8% in $\mathcal{J}$. Although the PASCAL-VOC dataset has higher variations than the Internet and iCoseg datasets, high-performance gains over the competing

Figure 3.3: The co-segmentation results generated by our approach on the iCoseg dataset. In the six examples (rows), the common object categories are Stonehenge, pyramids, pandas, kite-kitekid, and track and field, respectively.

methods can be obtained by our method. The results indicate that our method adapts itself well to unseen images with significant variations. Some examples of the co-segmentation results on the PASCAL-VOC dataset are shown in Figure 3.4. Compared with the Internet dataset in Figure 3.2 and the iCoseg dataset Figure 3.3, images on this dataset contain higher intra-class variations and subtle figure-ground differences. Our method can infer the common object segments of high quality. For example, the birds in the first row are of different colors and have a subtle figure-ground difference. It is difficult for hand-crafted features to handle this case well. Nevertheless, our method gets the promising segmentation results to owe to its ability of adaptive feature learning. Although our method does not adopt multi-scale learning which may make the running time longer and consume more memory, it still finds objects of different scales, such as those of object classes bus, dog, sofa, and train, because CNNs can tolerate scale variations to some extent.

Table 3.3: The performance of object co-segmentation on the PASCAL-VOC dataset under Jaccard index and Precision. The class-wise results are measured in Jaccard index. The numbers in red and green respectively indicate the best and the second best results.

Method	Avg. $\mathcal{P}$	Avg. $\mathcal{J}$	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
[31]	84.0	0.46	0.65	0.14	0.49	0.47	0.44	0.61	0.55	0.49	0.20	0.59	0.22	0.39	0.52	0.51	0.31	0.27	0.51	0.32	0.55	0.35
[86]	69.8	0.33	0.50	0.15	0.29	0.37	0.27	0.55	0.35	0.34	0.13	0.40	0.10	0.37	0.49	0.44	0.24	0.21	0.51	0.3	0.42	0.16
[13]	82.4	0.29	0.48	0.09	0.32	0.32	0.21	0.34	0.42	0.35	0.13	0.50	0.06	0.22	0.37	0.39	0.19	0.17	0.41	0.21	0.41	0.18
[116]	89.0	0.52	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
[43]	72.5	0.25	0.44	0.13	0.26	0.31	0.28	0.33	0.26	0.29	0.14	0.24	0.11	0.27	0.23	0.22	0.18	0.17	0.33	0.27	0.26	0.25
[66]	85.2	0.45	0.64	0.20	0.54	0.48	0.42	0.64	0.55	0.57	0.21	0.61	0.19	0.49	0.57	0.50	0.34	0.28	0.53	0.39	0.56	0.38
[65]	80.1	0.40	0.53	0.14	0.47	0.43	0.42	0.62	0.50	0.49	0.20	0.56	0.13	0.38	0.50	0.45	0.29	0.26	0.40	0.37	0.51	0.37
[142]	84.3	0.52	0.75	0.26	0.53	0.59	0.51	0.70	0.59	0.70	0.35	0.63	0.26	0.56	0.63	0 59	0.35	0.28	0.67	0.52	0.52	0.48
Ours	91.0	0.60	0.77	0.27	0.70	0.61	0.58	0.79	0.76	0.79	0.29	0.75	0.28	0.63	0.66	0.65	0.37	0.42	0.75	0.67	0.68	0.51

Ablation studies. In Table 3.1, w/o ℓ_m indicates the variant of our method where the mask loss ℓ_m is turned off, i.e., $\lambda = 0$ in Eq. (3.1). A performance drop about 2% is observed, but it still outperforms the state-of-the-arts. Therefore, the effectiveness of our method is mainly attributed to the proposed co-attention loss, instead of the mask loss with its adopted object proposals. We visualize the effect of using the mask loss in Figure 3.5. When the mask loss is turned off, the co-attention maps have many false positives and do not sharply cover the common objects. These co-attention maps result in the sub-optimal co-segmentation results. With the mask loss, the generator can highlight whole objects and suppress co-attention values in the background. Therefore, the attention maps result in much better co-segmentation results.

Our method employs *dense CRFs* [83] for post-processing and generating the co-segmentation results. To measure the effect of using dense CRFs in our method, we evaluate the performance of our method by replacing dense CRFs with Otsu's method and GrabCut [121] for post-processing. Otsu'method and GrabCut were adopted as the post-processing step to generate the co-segmentation results in previous work [66, 31, 116]. The results in Table 3.4 demonstrate that our method works well with each of the three schemes for post-processing.

In addition, our method can run with reasonable efficiency. Given 100 images for co-segmentation, model training and object mask refinement in each of 60 epochs take about 20 and 6 seconds, respectively, on an NVIDIA Titan X graphics card. Namely, it takes about $1,560$ seconds to estimate the co-segmentation results of 100 images, and the average time per image is 15.6 seconds.

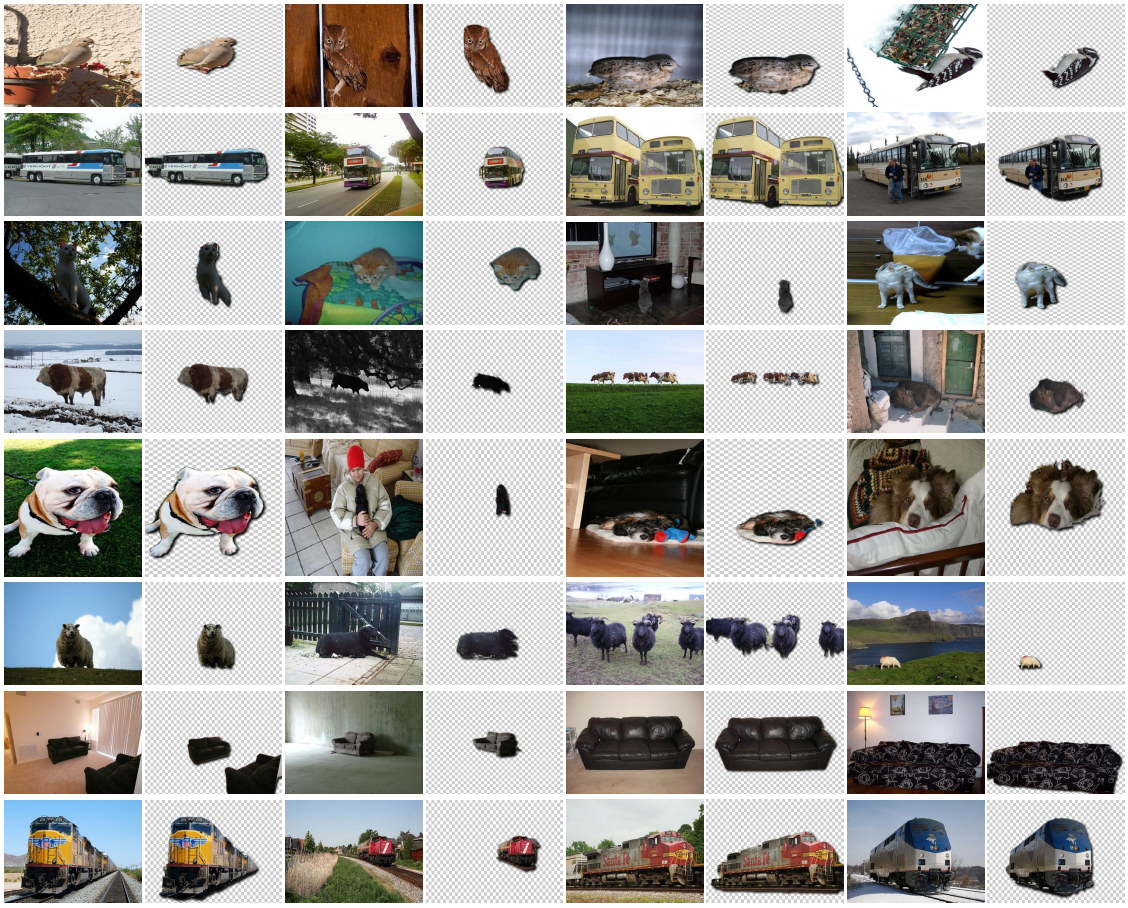


Figure 3.4: The co-segment results generated by our approach on the PASCAL-VOC dataset. From the first row to the last row, the classes are bird, bus, cat, cow, dog, sheep, sofa, and train, respectively.

3.3.3 Comparison with WSS methods

The setting of weakly supervised semantic segmentation (WSS) is similar to that of co-segmentation in the sense that images of specific categories are given for segmentation. Therefore, we compare our method with three state-of-the-art WSS methods, including [81, 72, 16] in Table 3.5. Note that we follow the previous methods for co-segmentation, i.e., those in Table 3.3, and use the PASCAL-VOC dataset collected in [31] as the test bed,

Table 3.4: The performance of our approach with three schemes for post-processing.

Method	Internet		iCoseg		PASCAL-VOC	
	$\mathcal{P}$	$\mathcal{J}$	$\mathcal{P}$	$\mathcal{J}$	$\mathcal{P}$	$\mathcal{J}$
Otsu's method	91.17	0.680	96.53	0.837	90.1	0.58
GrabCut	91.60	0.692	96.35	0.835	90.6	0.61
dense CRFs	92.29	0.698	96.46	0.835	91.0	0.60

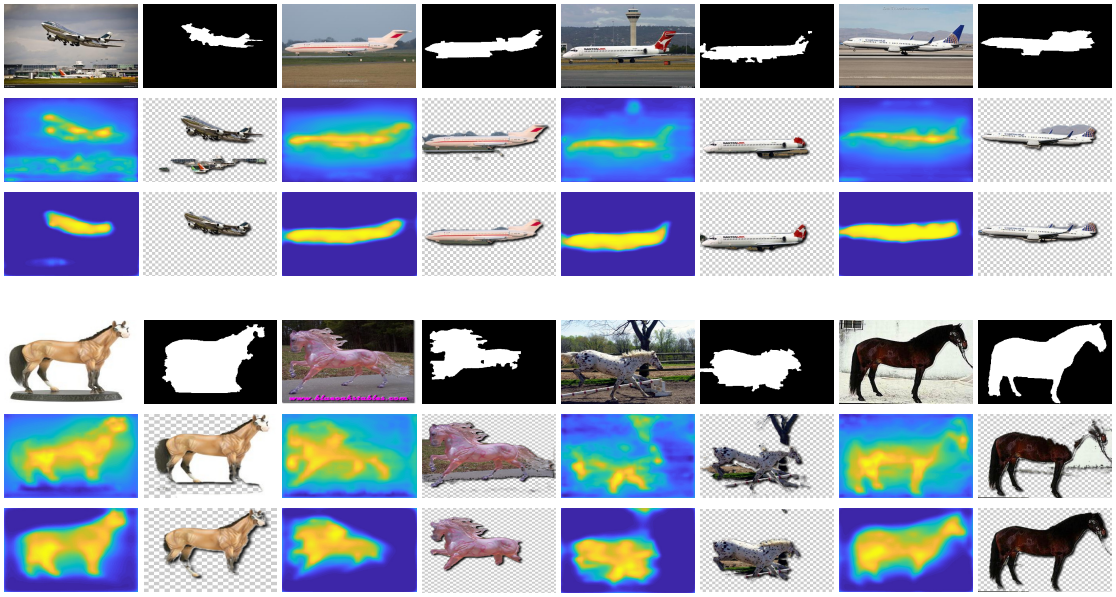


Figure 3.5: The effect of using the mask loss ℓ_m. The co-segmentation results on two object classes, including airplane (**top**) and horse (**bottom**). For each class, the first row shows four images and the corresponding estimated object masks, i.e., M_n in Eq. (3.7). When the mask loss ℓ_m is turned off, the second row gives the co-attention maps and the corresponding co-segmentation results. When the mask loss ℓ_m is turned on, the co-attention maps and the corresponding co-segmentation results displayed in the third row become better.

Table 3.5: The comparison of our method and three WSS methods on the PASCAL-VOC dataset under Jaccard index and Precision. The class-wise results are measured in Jaccard index. The numbers in red and green respectively indicate the best and the second best results.

Method	Avg. $\mathcal{P}$	Avg. $\mathcal{J}$.	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
[81]	90.4	0.57	0.68	0.28	0.61	0.41	0.62	0.79	0.67	0.71	0.32	0.67	0.31	0.65	0.60	0.66	0.53	0.44	0.68	0.65	0.49	0.57
[72]	89.0	0.56	0.71	0.29	0.60	0.55	0.57	0.74	0.71	0.76	0.21	0.80	0.15	0.72	0.74	0.66	0.52	0.44	0.80	0.41	0.49	0.43
[16]	92.0	0.59	0.78	0.29	0.64	0.63	0.59	0.82	0.74	0.68	0.31	0.75	0.21	0.63	0.67	0.66	0.49	0.34	0.74	0.62	0.70	0.53
Ours	91.0	0.60	0.77	0.27	0.70	0.61	0.58	0.79	0.76	0.79	0.29	0.75	0.28	0.63	0.66	0.65	0.37	0.42	0.75	0.67	0.68	0.51

instead of the standard PASCAL-VOC 2012 dataset [30], for evaluating the performance of object co-segmentation. In Table 3.5, our method outperforms the methods in [81, 72] and achieves a similar performance to that in [16]. Nevertheless, our method has two advantages over these WSS methods. First, our method does not require a training phase and does not rely on any background information. Second, our method can be applied to images of an arbitrary and unknown category. On the contrary, the models learned by WSS methods can segment only objects whose categories are covered in the training data.

Chapter 4

Graphical CNNs for Object Co-Saliency Detection

In this chapter[1], we proposed a CNN-based method for joint adaptive feature learning and co-saliency detection for given images without the pixel-wise annotation requirement. In the co-saliency detection, the detected object must be salient and repetitively appear in the given image set, so we decompose the object co-saliency detection into two complementary parts, *single-image saliency detection* and *cross-image co-occurrence region discovery*. In the former, we only detect the saliency object residing in a single image, which may not repetitively appear across images. In the latter, the regions repetitively appearing across images are discovers , which may not be visually salient. For this purpose, we proposed two novel losses, *the single-image saliency (SIS) loss* and *the co-occurrence (COOC) loss*, to capture the two different but complementary sour of information. These two losses measure the quality of the saliency maps by referring to individual images and the co-occurrence regions for each image pair, respectively. They are further integrated on a graphical model whose unary and pairwise terms correspond to the proposed SIS and COOC losses respectively, as illustrated in Figure 1.3 (a). Through optimizing the proposed losses, our approach can generate co-saliency maps of high quality by integrating SIS and COOC cues, as shown in Figure 1.3 (b). In the experiments, our proposed method outperforms the conventional unsupervised methods using engineered

[1]Published papers: [60]

features [66, 35, 105, 138, 137] and those using DL-based features [163, 161] because of joint adaptive feature learning and co-saliency detection based on CNNs. Compared with the conventional learning-based methods [39, 160], our method can achieve comparable or even slightly better performance and does not suffer from the high annotation cost of labeling object masks as training data. The proposed approach is comprehensively evaluated on three benchmark dataset, including *the MSRC dataset* [153], *the iCoseg dataset* [4], and *the Cosal2015 dataset* [161] and remarkably outperform the state-of-the-art method including the conventional methods and CNN-based methods.

In the following sections, we first give the literature review in Section 4.1. Next, we describe the proposed method in Section 4.2. Finally, the experiments are presented in Section 4.3.

4.1 Related work

4.1.1 Single-image saliency detection

Single-image saliency detection is to distinguish salient objects from the background by either unsupervised [154, 63, 71, 165, 139, 64] or supervised [144, 101, 51, 167, 166] methods based on color appearance, spatial locations, as well as various supplementary higher-level priors, including objectness. These approaches can handle well images with single salient objects. However, they may fail when the scenes are more complex, for example when multiple salient objects are presented with intra-image variations. By exploiting co-occurrence patterns when common objects are appearing in various images, co-saliency detection is expected to perform better. However, the appearance variations of common objects across images could also make co-saliency detection a more challenging task.

4.1.2 Co-saliency detection

Co-saliency detection discovers common and salient objects across multiple images using different strategies. The co-saliency detection methods have been developed within

the bottom-up frameworks based on different robust features, including low-level hand-crafted features [66, 35, 105, 138, 137, 160, 10, 96] and high-level DL-based semantic features [163, 161] to catch intra-image visual stimulus as well as inter-image repetitiveness. However, there are no features adopted suitable for all visual variations, and they treat the separate steps of feature extraction and co-saliency detection, leading to suboptimal performance.

Data-driven approaches [39, 148, 160] directly learn the patterns of co-salient objects to overcome the limitation of bottom-up methods. For instance, the transfer-learning-based method [160] uses the object masks to train a stacked denoising autoencoder (SDAE) to learn the intra-image contrast evidence, and propagate this knowledge to catch inter-image coherent foreground representations. Despite their impressive results, the performance might drop dramatically once the transferred knowledge on feature representations is not satisfactory as the separation of feature extraction and co-saliency detection may potentially impede the performance. Recently, Wei et al. [148] and Han et al. [39] have proposed unified learning-based methods to learn semantic features and detect co-salient objects jointly. Despite the improved performance, their methods rely on a large number of training object masks. It reduces the generalizability of their approaches to unseen images. However, our method can perform the adaptive and unified learning for given images in an unsupervised manner, and hence no issues as mentioned above exist in our approach.

4.1.3 Graphical models with CNNs

Deep learning has demonstrated success in many computer vision applications. For better preserving spatial consistency, graphical models have been integrated with CNNs when requiring structured outputs, such as depth estimation [99], stereo matching [79], semantic segmentation [11, 127, 168], image denoising, and optical flow estimation [147]. Although showing promise in preserving spatial consistency and modeling pairwise relationships, these methods have three major limitations when extending to co-saliency detection. First, their graphical models are built on single images, and hence cannot be directly ap-

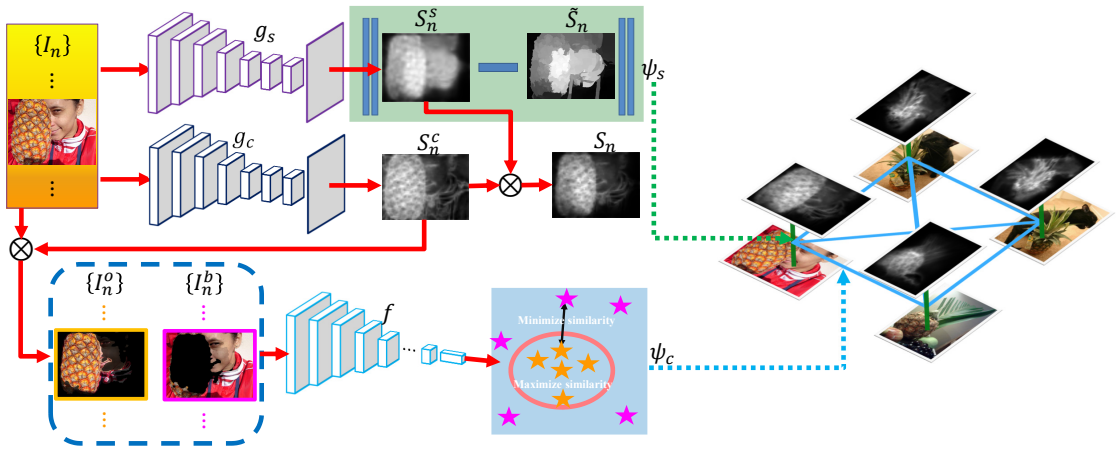


Figure 4.1: Overview of our approach to co-saliency detection. It optimizes an objective function defined on a graph by learning two collaborative FCN models g_s and g_c which respectively generates single-image saliency maps and cross-image co-occurrence maps.

plied to co-saliency detection with multiple images. Second, the pairwise terms in these graphical models often act as regularization terms to ensure spatial consistency but cannot work alone by themselves. Finally, they require training data to train the model. For the inter-image graphical models, Hayder et al. [44] and Yuan et al. [158] respectively integrated fully-connected CRFs into CNNs for object proposal co-generation and object co-segmentation, where each node is an object proposal. However, their methods still suffer from the last two limitations. In comparison, our approach integrates the merits from graphical models for co-saliency detection without the issues as mentioned earlier.

4.2 Proposed approach

We first describe the proposed formulation for co-saliency detection. Next, we offer a couple of enhancements by self-paced learning and fully connected conditional random fields. Finally, the optimization process and the implementation details are provided.

4.2.1 Proposed formulation

Given a set of N images $\{I_n\}_{n=1}^N$, co-saliency detection aims to detect the salient objects of a category commonly present in these images. We accomplish the task by decomposing it into two sub-tasks, *single-image saliency detection* and *cross-image co-occurrence*

discovery. The former detects the salient regions in a single image, without considering whether the identified regions are commonly present across images. The latter discovers the regions repetitively occurring across images while disregarding whether the discovered regions stand out visually. Co-saliency detection, finding the salient co-occurrence regions, can then be carried out by performing and integrating the two tasks on a graph whose two types of edges respectively correspond to the two tasks, as shown in Figure 1.3 (a). The proposed objective function on the graph is defined by

$$E(\mathbf{w}) = \sum_{n=1}^{N} \psi_s(I_n; \mathbf{w}) + \sum_{n=1}^{N} \sum_{m \neq n} \psi_c(I_n, I_m; \mathbf{w}), \tag{4.1}$$

where the unary term $\psi_s(I_n; \mathbf{w})$ focuses on **s**aliency detection for the image I_n, the pairwise term $\psi_c(I_n, I_m; \mathbf{w})$ accounts for **c**o-occurrence discovery for the image pair (I_n, I_m), and $\mathbf{w}$ is the set of model parameters.

As shown in Figure 4.1, we learn two fully convolutional networks (FCNs) [106] models, g_s and g_c, to optimize the unary term ψ_s and the pairwise term ψ_c in Eq. (4.1), respectively. For image I_n, FCN g_s investigates intra-image clues and generates its *saliency map* S_n^s. In contrast, FCN g_c discovers cross-image evidence and produces its *co-occurrence map* S_n^c, where the repetitively occurring regions are highlighted. The resultant *co-saliency map*, highlighting the co-occurrence and salient regions, is yielded by $S_n = g_s(I_n) \otimes g_c(I_n) = S_n^s \otimes S_n^c$, where $\otimes$ denotes the element-wise multiplication operator.

Let $\mathbf{w}_s$ and $\mathbf{w}_c$ denote the learnable parameter sets of FCNs g_s and g_c, respectively. We learn g_s and g_c jointly by optimizing $E(\mathbf{w} = \mathbf{w}_s \cup \mathbf{w}_c)$ in Eq. (4.1). The unary term ψ_s and the pairwise term ψ_c in Eq. (4.1) are described below.

4.2.2 Unary term ψ_s

This term aims to identify the salient regions in a single image. It guides the training of FCN g_s, which produces saliency map S_n^s for image I_n, i.e., $S_n^s = g_s(I_n)$. Inspired by Zhang et al. [162], we apply an existing unsupervised method for saliency detection to image I_n and use its output saliency map $\tilde{S}_n$ as the desired target for learning FCN g_s. In

this work, we adopt MILP [64] to generate $\tilde{S}_n$. Specifically, the unary term $\psi_s(I_n; \mathbf{w}_s)$ applied to image I_n is defined by

$$\psi_s(I_n; \mathbf{w}_s) = \sum_{i \in I_n} R_n(i) |S_n^s(i) - \tilde{S}_n(i)|^2, \tag{4.2}$$

where i is the index of the pixels in I_n, $S_n^s(i)$ and $\tilde{S}_n(i)$ are respectively the saliency values of maps S_n^s and $\tilde{S}_n$ at pixel i, and $R_n(i)$ represents the importance of pixel i. Pixels in map $\tilde{S}_n$ can be divided into the salient and non-salient groups by using the mean value of $\tilde{S}_n$ as the threshold. $R_n(i)$ is introduced here to deal with the potential size unbalance between the two groups. Let δ be the ratio of salient pixels over the whole image I_n. $R_n(i)$ takes the value $1 - \delta$ if pixel i belongs to the salient group, and δ otherwise. In this way, the salient and non-salient groups contribute equally in Eq. (4.2).

4.2.3 Pairwise term ψ_c

The pairwise term ψ_c seeks the regions simultaneously appearing across images. It serves as the objective to learn FCN g_c. The regions should look similar across images but distinctive from surrounding non-detected regions. Thus, two criteria are jointly considered in the design of ψ_c, including 1) high cross-image similarity between the detected co-occurrence regions and 2) high intra-image distinctness between the detected co-occurrence regions and the rest of the image. The second criterion is auxiliary but crucial to avoid trivial solutions.

As shown in Figure 4.1, FCN g_c produces the co-occurrence map S_n^c for image I_n, i.e., $S_n^c = g_c(I_n)$. The sigmoid function is used as the activation function in the last layer of g_c. Thus, the value of the co-occurrence map at each pixel i, $S_n^c(i)$, is between 0 and 1. With S_n^c, image I_n is decomposed into two masked images,

$$I_n^o = S_n^c \otimes I_n \;\text{ and }\; I_n^b = (1 - S_n^c) \otimes I_n, \tag{4.3}$$

where $\otimes$ denotes element-wise multiplication. The masked image I_n^o keeps the detected co-occurrence regions of I_n, while image I_n^b contains the rest.

To measure the similarity between images, we employ a feature extractor f to compute the features of a given image. In this work, the extractor f can be a pre-trained CNN model for image classification, e.g., AlexNet [85] or VGG-19 [131], with the softmax function and the last fully connected layer removed. We apply the extractor f to all masked images $\{I_n^o, I_n^b\}_{n=1}^N$ and obtain their features $\{f(I_n^o) \in \mathbb{R}^c, f(I_n^b) \in \mathbb{R}^c\}_{n=1}^N$, where c is the feature dimension. With these extracted features, the pairwise term $\psi_c(I_n, I_m; \mathbf{w}_c)$ applied the image pair I_n and I_m is defined by

$$\psi_c(I_n, I_m; \mathbf{w}_c) = -\log(p_{nm}), \tag{4.4}$$

where p_{nm} is the score estimating the quality of the detected co-occurrence regions in I_n and I_m. The score p_{nm} is defined below,

$$p_{nm} = \frac{\exp(-d_{nm}^+)}{\exp(-d_{nm}^+) + \exp(-d_{nm}^-)}, \text{ where} \tag{4.5}$$

$$d_{nm}^+ = \frac{1}{c}\|f(I_n^o) - f(I_m^o)\|^2 \text{ and} \tag{4.6}$$

$$d_{nm}^- = \frac{1}{2c}\|f(I_n^o) - f(I_n^b)\|^2 + \frac{1}{2c}\|f(I_m^o) - f(I_m^b)\|^2. \tag{4.7}$$

Eq. (4.6) measures the inter-image distance between the detected co-occurrence regions in images I_n and I_m (criterion 1). Eq. (4.7) evaluates the intra-image distance between the detected co-occurrence regions and the rest of the image (criterion 2). By minimizing the pairwise term $\psi_c(I_n, I_m; \mathbf{w}_c)$ in Eq. (4.4) for each image pair (I_n, I_m), the resultant FCN g_c will produce the co-occurrence maps where the inter-image distances between the detected co-occurrence regions are minimized while the intra-image distances between the detected co-occurrence regions and the rest of the images are maximized. After learning FCNs g_s and g_c jointly through the unary and pairwise terms in Eq. (4.1), the resultant co-saliency map S_n of a given image I_n is produced via $S_n = g_s(I_n) \otimes g_c(I_n)$.

Note that the pairwise term in Eq. (4.4) is defined by referring to the co-occurrence maps produced by FCN g_c, i.e., S_n^c and S_m^c. In practice, we found that the performance of co-saliency detection can be further improved if co-saliency maps S_n and S_m are also

taking into account in the pairwise term. In our implementation, we extend the pairwise term in Eq. (4.4) to

$$\psi_c(I_n, I_m; \mathbf{w}_c) = -\lambda_c \log(p_{nm}) - \lambda_{\tilde{c}} \log(\tilde{p}_{nm}), \tag{4.8}$$

where like p_{nm}, $\tilde{p}_{nm}$ is computed in the same way but by referring to co-saliency maps S_n and S_m. Constants λ_c and $\lambda_{\tilde{c}}$ are used in Eq. (4.8) for weighting the corresponding terms. In the following, we will show that the quality of the co-saliency maps can be further improved via two extensions, including map enhancement by self-paced learning and post-processing by fully connected conditional random fields (or DenseCRFs) [83].

4.2.4 Co-saliency map enhancement

The self-paced learning with CNNs is proposed to make salience map sharper. Then, fully connected conditional random fields are adopted to preserve co-salient objects' boundaries. The details of these two extensions are given below.

Co-saliency map enhancement by self-paced learning. The co-saliency maps obtained by optimizing Eq. (4.1) are sometimes over smooth, because both FCNs g_s and g_c do not take into account the information regarding object boundaries. To address this issue, we oversegment each image I_n into superpixels $Q_n = \{q_n^k\}_{k=1}^K$, where q_n^k is the kth superpixel and K is the number of superpixels. Pixels in a superpixel tend to belong to either a salient object or the background together. This property can be leveraged to propagate information from pixels of high confidence to those of low confidence within the same superpixel. We divide superpixels into three groups, i.e., $Q_n = O_n \cup B_n \cup T_n$. The first two groups, O_n and B_n, contain superpixels that likely belong to the object and the background, respectively. The third group T_n covers the rest. Given the co-saliency map S_n,

the three groups are yielded by

$$q_n^k \in \begin{cases} O_n, & \text{if } \mu_n^k > \mu_n + \sigma_n, \\ B_n, & \text{if } \mu_n^k < \mu_n - 0.25 * \sigma_n, \\ T_n, & \text{otherwise,} \end{cases} \text{ for } k = 1, 2, ..., K, \tag{4.9}$$

where μ_n^k is the mean saliency value of superpixel q_n^k, while μ_n and σ_n are the mean and the standard deviation of $\{\mu_n^k\}_{k=1}^K$. Besides, we follow the background seed sampling strategy used in previous work [154, 68], and add superpixels on the image boundary to the set B_n. Superpixels in O_n and B_n are confident to be assigned to either the salient regions or the background. Those in T_n are ambiguous, so they are not taken into account here. With O_n and B_n of image I_n, another FCN g_e for co-saliency map **e**nhancement is trained by optimizing

$$\psi_e(I_n; \mathbf{w}_e) = w_o \sum_{q \in O_n} \sum_{i \in q} \|S_n^e(i) - 1\|^2 + w_b \sum_{q \in B_n} \sum_{i \in q} \|S_n^e(i) - 0\|^2, \tag{4.10}$$

where FCN g_e generates map $S_n^e = g_e(I_n)$, and i the index of pixels in I_n. Constants $w_o = \frac{|B_n|}{|O_n|+|B_n|}$ and $w_b = \frac{|O_n|}{|O_n|+|B_n|}$ are the weights used to balance the contributions of O_n and B_n, where $|O_n|$ and $|B_n|$ are the numbers of pixels in O_n and B_n, respectively.

The term in Eq. (4.10) enhances the consensus within superpixels of high confidence. If it is turned on, the objective is extended from that in Eq. (4.1) to

$$E(\mathbf{w}) = \sum_{n=1}^{N} \psi_s(I_n; \mathbf{w}_s) + \lambda_e \psi_e(I_n; \mathbf{w}_e) + \sum_{n=1}^{N-1} \sum_{m=n+1}^{N} \psi_c(I_n, I_m; \mathbf{w}_c), \tag{4.11}$$

where λ_e is a weight, and $\mathbf{w} = \mathbf{w}_s \cup \mathbf{w}_e \cup \mathbf{w}_c$ is the union of the learnable parameter sets of FCNs g_s, g_e, and g_c. After optimizing the objective function in Eq. (4.11), the co-saliency map S_n of image I_n is generated by $S_n = g_s(I_n) \otimes g_e(I_n) \otimes g_c(I_n) = S_n^s \otimes S_n^e \otimes S_n^c$.

Post-processing using DenseCRFs. The co-saliency maps obtained by optimizing the objective in Eq. (4.11) can be further improved by enforcing spatial coherence and pre-

serving object boundaries. To this end, we follow previous work [89, 51] and adopt DenseCRFs [83] to postprocess the co-saliency map S_n of a given image I_n. We use the DenseCRFs code implemented by Li and Yu [89] in this work.

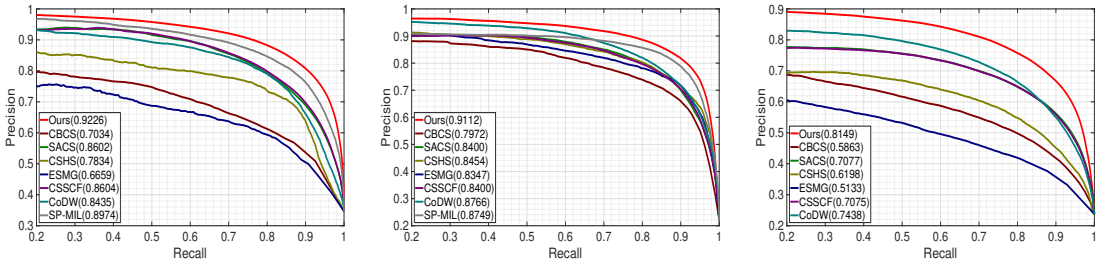
4.2.5 Optimization

To reduce memory consumption and speed up the training, the proposed method is optimized by using a two-stage procedure. At the first stage, we respectively learn FCNs g_s and g_c by using the objective functions in Eq. (4.2) and Eq. (4.4) with all images for 20 epochs. The co-saliency maps $\{S_n = g_s(I_n) \otimes g_c(I_n)\}_{n=1}^{N}$ become stable enough. Thus, we divide the superpixels of each image into three groups via Eq. (4.9). FCN g_e is then trained with the objective in Eq. (4.10) with all images for 20 epochs. At the second stage, we turn on all the three terms in Eq. (4.11) where the extended pairwise term in Eq. (4.8) is adopted. The three FCNs, g_s, g_e, and g_c, are optimized jointly for 20 epochs. Note that at the second stage, we optimize only the parameters in the last two convolutional layers and the skip connections of each FCN model.

The objectives in Eq. (4.1) and Eq. (4.11) are defined on a fully-connected graph. It is difficult to directly optimize either objective with all images at the same time due to the limited memory size. Thereby, we adopt the *piecewise training* scheme [133]. Namely, we consider only the sub-graph yielded by a subset of images at each time. The learning rate is set to 10^{-6} at the first stage and is reduced to 10^{-8} at the second stage. The weight decay and momentum are set to 0.0005 and 0.9, respectively. The objective function in Eq. (4.11) is differentiable. We choose Adam [78] as the optimization solver for its rapid convergence. The gradients with respect to the optimization variables can be derived straightforward, so we omit their derivation here.

4.2.6 Implementation details

The proposed method is implemented using $\mathrm{MatConvNet}$ [141]. The same network architecture is used in all the experiments. ResNet-50 [47] is adopted as the feature extractor f for the pairwise term because AlexNet [85] and VGG-16/19 [131] sometimes lead to the



MSRC iCoseg Cosal2015

Figure 4.2: Comparison with the state-of-the-art methods with the same setting in terms of PR curves on three benchmark datasets. The numbers in parentheses are AP values.

problem of vanishing gradients in our application. The feature extractor f is pre-trained on ImageNet [29] and fixed during the optimization process. The features extracted by f are the inputs to the last fully connected layer of f. The feature dimension, i.e., c in Eq. (4.6) and Eq. (4.7), is set to $2{,}048$. All FCNs, including g_s, g_e and g_c, are developed based on the VGG-16 [131] setting of FCN [106]. We replace the activation function *softmax* in the last layer with the *sigmoid* function. SLIC [1] is adopted to generate superpixels because of its computational efficiency, better compactness and regularity. The models pre-trained on the ImageNet [29] dataset for classification are required. Following previous co-saliency detection methods [163, 161], we determine the values of hyperparameters empirically and keep them fixed in all the experiments.

4.3 Experimental results

In this section, we first describe the datasets and evaluation metrics. Next, we compare our method with a set of state-of-the-art methods. Finally, we investigate contributions of different components by reporting ablation studies.

4.3.1 Datasets and evaluation metrics

Datasets. We evaluated the proposed approach on three public benchmark datasets: *iCoseg* [4], *MSRC* [153] and *Cosal2015* [161]. *iCoseg* consists of 38 groups of total 643 images, and each group has $4 \sim 42$ images. The images of *iCoseg* contain single or multi-

Table 4.1: The performance of co-saliency detection on three benchmark datasets. SI and CS denote the single-image saliency and co-saliency methods, respectively. US and S indicate the unsupervised and supervised methods, respectively. The numbers in red and green respectively indicate the best and the second best results of the unsupervised co-saliency methods (CS+US), the group which the proposed method belongs to.

Method	Setting	MSRC			iCoseg			Cosal2015		
		AP	F_β	S_α	AP	F_β	S_α	AP	F_β	S_α
DIM [160]	CS+S	-	-	-	0.8773	0.7918	0.7583	-	-	-
UMLBF [39]	CS+S	0.9160	0.8410	-	-	-	-	0.8210	0.7120	-
CBCS [35]	CS+US	0.7034	0.5910	0.4801	0.7972	0.7408	0.6580	0.5863	0.5579	0.5439
SACS [10]	CS+US	0.8602	0.7877	0.7074	0.8400	0.7973	0.7523	0.7077	0.6923	0.6938
CSHS [105]	CS+US	0.7834	0.7118	0.6661	0.8454	0.7549	0.7502	0.6198	0.6181	0.5909
ESMG [96]	CS+US	0.6659	0.6245	0.5804	0.8347	0.7766	0.7677	0.5133	0.5114	0.5446
CSSCF [66]	CS+US	0.8604	0.8005	0.7383	0.8400	0.7811	0.7404	0.7075	0.6815	0.6710
CoDW [161]	CS+US	0.8435	0.7724	0.7129	0.8766	0.7985	0.7500	0.7438	0.7046	0.6473
SP-MIL [163]	CS+US	0.8974	0.8029	0.7687	0.8749	0.8143	0.7715	-	-	-
MVSRC [156]	CS+US	0.8530	0.7840	-	0.8680	0.8100	-	-	-	-
Ours	CS+US	0.9226	0.8404	0.7948	0.9112	0.8497	0.8200	0.8149	0.7580	0.7506
LEGS [144]	SI+S	0.8479	0.7701	0.6997	0.7924	0.7473	0.7529	0.7339	0.6926	0.7068
DCL [89]	SI+S	0.9065	0.8259	0.7742	0.9003	0.8444	0.8606	0.7815	0.7386	0.7591
DSS [51]	SI+S	0.8700	0.8313	0.7435	0.8802	0.8386	0.8483	0.7745	0.7509	0.7579
UCF [167]	SI+S	0.9217	0.8114	0.8175	0.9292	0.8261	0.8754	0.8081	0.7194	0.7790
Amulet [166]	SI+S	0.9219	0.8159	0.8162	0.9395	0.8381	0.8937	0.8201	0.7384	0.7856
GMR [154]	SI+US	0.8092	0.7460	0.6547	0.7990	0.7805	0.7068	0.6649	0.6605	0.6599
GP [71]	SI+US	0.8200	0.7422	0.6844	0.7821	0.7495	0.7198	0.6847	0.6576	0.6714
MB+ [165]	SI+US	0.8367	0.7817	0.7200	0.7868	0.7706	0.7272	0.6710	0.6689	0.6724
MST [139]	SI+US	0.8057	0.7491	0.6460	0.8019	0.7659	0.7292	0.7096	0.6669	0.6676
MILP [64]	SI+US	0.8334	0.7776	0.6871	0.8182	0.7883	0.7514	0.6797	0.6734	0.6752
SVFSal [162]	SI+US	0.8669	0.7934	0.7688	0.8376	0.8056	0.8271	0.7468	0.7120	0.7604

ple similar objects in various poses and sizes with complex backgrounds. *MSRC* contains 7 groups of total 240 images, and each group has $30 \sim 53$ images. Compared to *iCoseg*, objects in *MSRC* exhibit greater appearance variation. *Cosal2015* is a more recent and more challenging dataset than the other two. It has 50 groups and a total of 2015 images. Each group contains 26 to 52 images, with various poses and sizes, appearance variations and even more complex backgrounds. Because the images of *iCoseg* and *Cosal2015* have larger sizes than the ones of *MSRC*, different batch sizes and resolutions were used. The batch size is 3, and the resolution is 512×512 for *iCoseg* and *Cosal2015*, while the batch size is 5, and the resolution is 320×320 for *MSRC*.

Evaluation metrics. To evaluate the performance of co-saliency detection, we consider three metrics, *average precision* (AP), *F-measure* (F_β), and *structure measure* (S_α). AP is computed from the area under the Precision-Recall (PR) curve, which is produced by binarizing saliency maps with every integer threshold in the range of $[0, 255]$. *F-measure*

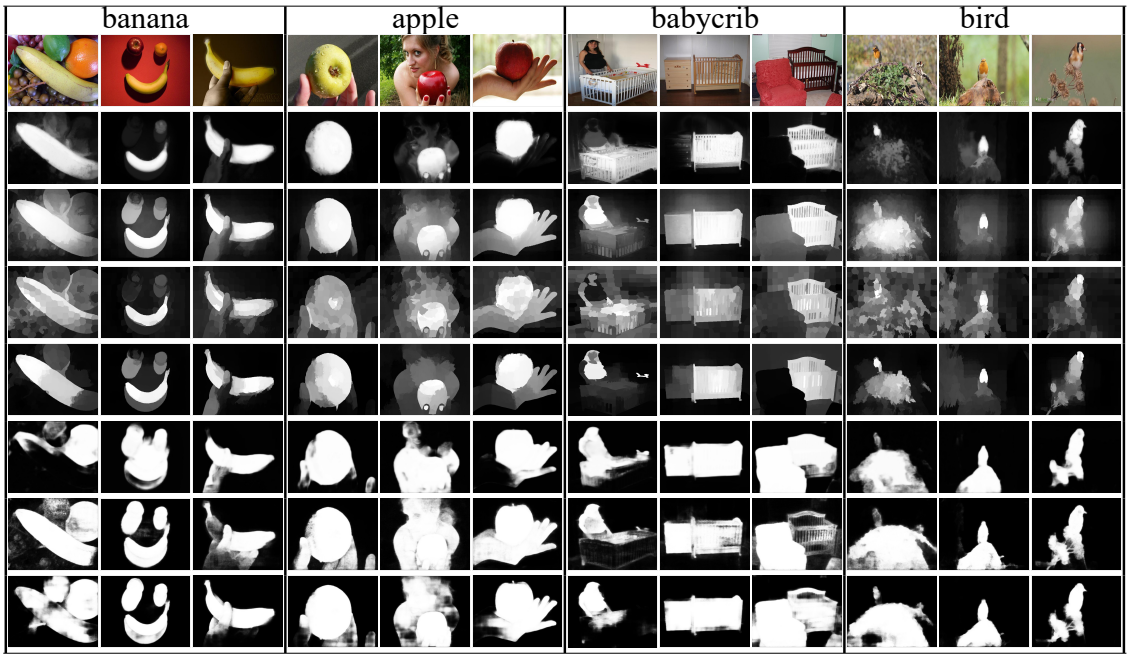


Figure 4.3: Example saliency maps generated by our method and some state-of-the-art methods. From the top to the bottom, they are the given images, ours, CSSCF [66], CoDW [161], MILP [64], SVFSal [162], UCF [167] and Amulet [166].

denotes the harmonic mean of the precision and recall values obtained by a self-adaptive threshold $T = \mu + \sigma$, where μ and σ are respectively the mean and standard deviation of the saliency map. With the precision and recall values, the *F-measure* is computed by $F_\beta = \frac{(1+\beta^2)\times precision \times recall}{\beta^2 \times precision + recall}$, where $\beta^2 = 0.3$ to emphasize more on recall as suggested in previous work [163, 161, 8]. The *structure measure* (S_α) [33] is adopted to evaluate the spatial structure similarities of saliency maps based on both region-aware structural similarity S_r and object-aware structural similarity S_o, defined as $S_\alpha = \alpha * S_r + (1-\alpha) * S_o$, where $\alpha = 0.5$ following [33].

4.3.2 Comparison with state-of-the-art methods

To have a thorough comparison with state-of-the-art methods, we divide them into four groups, i.e., the unsupervised saliency [154, 71, 165, 139, 64, 162] and co-saliency [66, 35, 105, 163, 161, 10, 96, 156] detection methods as well as supervised saliency [144, 89, 51, 167, 166] and co-saliency [39, 160] detection methods. The overall performance statistics are compared in Table 4.1 and Figure 4.2. Please note that all compared su-

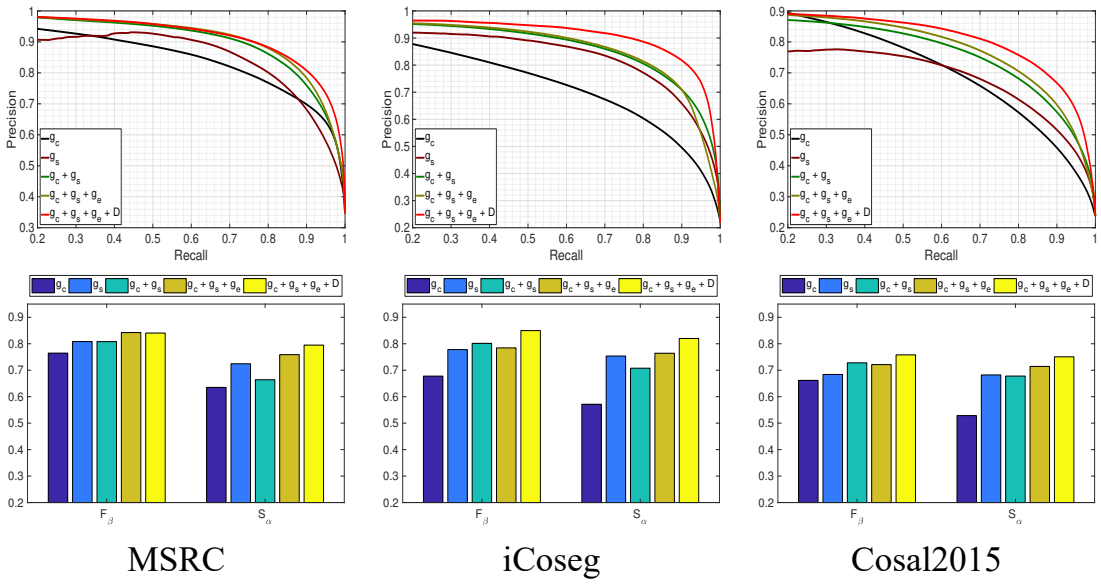


Figure 4.4: Ablation studies on three benchmarks. The top row plots the PR curves, while the bottom row shows the performance in F_β and S_α.

pervised single-image saliency detection methods are CNN-based. Among unsupervised single-image saliency methods, SVFSal [162] is CNN-based. When available, we used the publicly released source code with default parameters provided by the authors to reproduce the experimental results. For methods without releasing source code, we either evaluated metrics on their pre-generated co-saliency maps (SP-MIL [163], CoDW [161] and DIM [160]) or directly copied the numbers reported in their papers (UMLBF [39] and MVSRC [156]).

From Table 4.1, our method outperforms all methods with the same unsupervised co-saliency detection setting by a significant margin. Most approaches of this category take feature extraction and co-salient object detection as separating steps. Our approach excels them by performing these steps simultaneously and adopting CNN models. Comparing with the group of the supervised co-saliency method, UMLBF [39] and DIM [160], our method yields comparable or even slightly better performance without expensive object annotations. Although both with the unsupervised setting, by taking advantage of additional information within an image set, our method outperforms the group of unsupervised single-image saliency detection methods. It's worth mentioning that our approach also beats the unsupervised CNN-based single-saliency method, SVFSal [162] that requires

Figure 4.5: Example co-saliency maps generated by combinations of different components. From the top to the bottom, they are the given images, g_c, g_s, g_c+g_s, g_c+g_s+g_e and g_c+g_s+g_e+D, respectively.

saliency proposal fusion for generating high-quality pseudo-ground-truth as training data. In general, the supervised CNN-based single-image saliency methods perform the best among four groups of methods as they better utilize the object annotations. Even so, our method still outperforms many of the methods in this group by exploiting cross-image referencing and adaptive feature learning. From the PR curves in Figure 4.2, the proposed method exceeds the state-of-the-art approaches by a large margin.

Figure 4.3 shows example saliency maps produced by our method and some state-of-the-art methods, including unsupervised co-saliency detection methods (CSSCF [66], CoDW [161]), unsupervised single-image saliency methods (MILP [64] and SVFSal [162]), and supervised CNN-based methods (UCF [167] and Amulet [166]). Without referring to other images in the given image set, single-image saliency methods could detect the visually salient objects that do not repetitively appear in other images, such as the orange and the apple in the second image of the banana set or the woman in the first image of the babycrib set. Co-saliency detection methods perform better in this regard. The competing co-saliency methods, CSSCF [66] and CoDW [161], cannot perform well for images with low figure-ground discrepancies or highly-textured backgrounds, such as the second and third images of the babycrib set or the first and second images of the bird set. The major drawback of their approaches is to treat feature extraction as a separate step. Thus,

they cannot find the most discriminative features across images. Our method addresses the problem by performing adaptive feature learning and co-saliency detection jointly.

4.3.3 Ablation studies

We have performed ablation studies to investigate the contributions of individual components, g_c, g_s, g_e, and DenseCRFs. Figure 4.4 reports the results with different metrics. $+$D denotes the results refined by DenseCRFs. For both AP and F_β, the integration of g_c and g_s outperforms either alone. It is not the case for S_α measuring the structure of the detected objects. Both self-paced learning and DenseCRFs further improve the results.

Figure 4.5 gives the example co-saliency maps for ablation studies. They demonstrate that g_c and g_s can be complementary to each other. Taking the butterfly set as an example, g_s highlights both butterflies and flowers in the first, third and fourth images. After integrating the co-occurrence information discovered by g_c, the flowers are mostly removed and lightened in g_c+g_s. As mentioned above, g_c+g_s could perform worse in terms of S_α. It is because g_c tends to have less certainty, particularly inside objects or ambiguous background regions, as illustrated in the second row of Figure 4.5. Thus, g_c+g_s usually generates fuzzier maps than g_s alone. For example, the cattle have lower saliency values in g_c+g_s (the fourth row of Figure 4.5) than g_s (the third row of Figure 4.5). By propagating information from regions with high confidence, g_e improves the certainty of the results of g_c+g_s. Although with less gain in AP and F_β, it brings significant improvement in S_α since objects are more highlighted and the backgrounds are further lightened as shown in the fifth row of Figure 4.5. Finally, the DenseCRF enhances spatial coherence and boundary preservation, thus improving both quantitative and qualitative results.

Chapter 5

CNN-based Instance-Level Object Co-Segmentation

In this chapter[1], we develop a CNN-based method for instance co-segmentation. Based on the problem setting, our method has no access to annotated instance masks for learning and cannot involve any pre-training process. Inspired by Zhou et al. [171]'s observation that object instances often cover the *peaks* in a response map of a classier, we design a novel *co-peak* loss to detect the common peaks (or co-peaks for short) in two images. The co-peak loss is built upon a 4D tensor that is learned to encode the inter-image similarity at every location. The co-peaks inferred from the learned 4D tensor correspond to two locations, one in each of the two images, where discriminative and similar features are present. Therefore, the two locations are potentially covered by two object instances. Using the co-peak loss alone may lead to unfavorable false positives and negatives. Thus, we develop the *affinity* loss and the *saliency* loss to complement the co-peak loss. The former carries out discriminative feature learning for the 4D tensor construction by separating the foreground and background features. The latter estimates the co-saliency maps to localize the co-salient objects in an image, and can make our model focus on co-peak search in co-salient regions. The three loss functions work jointly and can detect co-peaks of high quality. We design a ranking function taking the detected co-peaks and co-saliency maps as inputs and accomplish instance mask segmentation by selecting object proposals.

[1]Published papers: [58]

In the experiments, because this task is new, the four datasets are collected from three standard instance segmentation datasets including the MS COCO [98], PASCAL VOC 2012 [30, 41], and SOC [32] datasets. We extensively evaluate the proposed method and its variant for object co-localization on these four collected datasets, and our method performs favorably against the state-of-the-art methods.

In the following sections, we first give the literature review in Section 5.1. Next, we describe the proposed method in Section 5.2. Finally, the experiments are presented in Section 5.3.

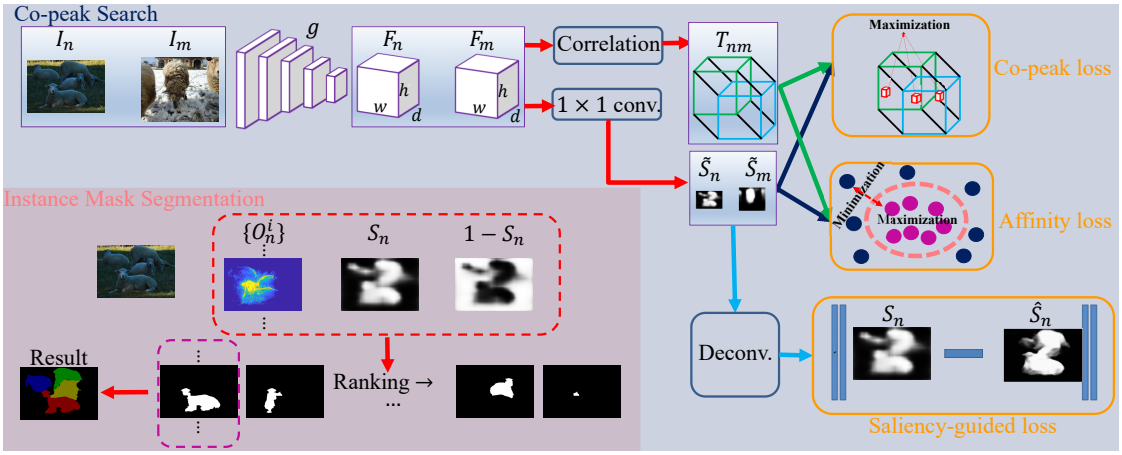


Figure 5.1: Overview of our method, which contains two stages, *co-peak search* within the blue-shaded background and *instance mask segmentation* within the red-shaded background. For searching co-peaks in a pair of images, our model extracts image features, estimates their co-saliency maps, and performs feature correlation for co-peak localization. The model is optimized by three losses, including the co-peak loss ℓ_t, the affinity loss ℓ_a, and the saliency loss ℓ_s. For instance mask segmentation, we design a ranking function taking the detected co-peaks, the co-saliency maps, and the object proposals as inputs, and select the top-ranked proposal for each detected instance.

5.1 Related work

5.1.1 Object co-segmentation

This task [40, 66, 116, 124, 136, 142] aims to segment the common objects in images. Its major difficulties lie in large intra-class variations and background clutter. Most methods rely on robust features, such as handcrafted and deep learning based features, for address-

ing these difficulties. In addition, saliency evidence, including single-image saliency [36, 65, 66, 124, 135, 56] or multi-image co-saliency [15, 136], has been explored to localize the salient and common objects. Recently, CNN-based methods [57, 92, 158] achieve better performance by joint representation learning and co-segmentation.

Despite effectiveness, the aforementioned methods do not provide instance-level results. In this work, we go beyond object co-segmentation and investigate instance co-segmentation. Our method can determine the number, locations, and contours of common instances in each image, and offers instance-aware image understanding.

5.1.2 Object co-localization

This task [22, 26, 134, 149, 150] discovers the common instances in images. Different from object co-segmentation, it is instance-aware. It detects and outputs the bounding box of a single instance in each image even if multiple instances are present in the image. Compared with object co-localization, instance co-segmentation identifies all instances in an image in the form of instance segments.

5.1.3 Instance-aware segmentation

Instance-aware segmentation includes *class-aware* [3, 27, 42, 46, 171] and *class-agnostic* [34, 61, 88] methods. Given training data of predefined categories, class-aware instance segmentation, aka instance segmentation, learns a model to seek each object instance belonging to one of these categories. A widely used way for instance segmentation is to first detect instance bounding boxes and then segment the instances within the bounding boxes [27, 42, 45, 46, 95, 103, 112]. Another way is to directly segment each instance without bounding box detection [3, 82, 97, 104, 171]. While most methods for instance segmentation are supervised, Zhou et al. [171] present a weakly supervised one. All these methods for instance segmentation rely on training data to learn the models. Despite the effectiveness and efficiency in testing, their learned models are not applicable to unseen object categories.

In practice, it is difficult to enumerate all object categories of interest in advance

and prepare class-specific training data, which limits the applicability of class-aware instance segmentation. Class-agnostic instance segmentation [34, 61, 88] aims at segmenting object instances of arbitrary categories, and has drawn recent attention. It is challenging because it involves both generic object detection and segmentation. Instance co-segmentation is highly related to class-agnostic instance segmentation in the sense that both of them can be applied to arbitrary and even unseen object categories. However, existing class-agnostic methods require annotated training data in the form of object contours. On the contrary, our method for instance co-segmentation explores the mutual information regarding the common instances in given images, and does not need any pre-training procedure on additional data annotations. Thus, our method has better generalization.

5.2 Proposed approach

In this section, we give an overview of our method, describe its components, *co-peak search* and *instance mask segmentation*, and provide the implementation details.

5.2.1 Overview

Suppose that a set of images $D = \{I_n\}_{n=1}^{N}$ consisting of object instances of a particular category is given, where $I_n \in \mathbb{R}^{W \times H \times c}$ is the nth image while W, H, and c are the width, the height, and the number of channels of I_n, respectively. The goal of instance co-segmentation is to identify and segment each of all instances in D. Note that no training data with pixel-wise annotations are provided. In addition, both the object category and the number of instances in each image are unknown.

In the proposed method, we decompose instance co-segmentation into two stages, i.e., *co-peak search* and *instance mask segmentation*. The overview of our method is shown in Figure 5.1, where the two stages are highlighted with the blue-shaded area and the red-shaded backgrounds, respectively.

At the stage of co-peak search, we aim to seek co-peaks in the response maps of two

images, where a co-peak corresponds two discriminative and similar points, one in each image, so that each point is potentially within an object instance. We design a network model for co-peak detection. The front part of our model is a fully convolutional network (FCN) g, which extracts the feature maps of input images. After feature extraction, our model is split into two streams. One stream correlates the feature maps of two images for co-peak localization. The other estimates the co-saliency maps of input images, which in turn enforces FCN g to generate more discriminative feature maps. Our model is optimized by three novel losses, including the co-peak loss ℓ_t, the affinity loss ℓ_a, and the saliency loss ℓ_s. After optimization, co-peaks are detected and co-saliency maps are estimated. At the stage of instance mask segmentation, we design a ranking function that takes the detected co-peaks, the estimated co-saliency maps, and the instance proposals into account, and yield one mask for each detected instance.

5.2.2 Co-peak search

As shown in Figure 5.1, our model takes a pair of images, I_n and I_m, from D as input at a time. It first extracts the feature maps $F_n \in \mathbb{R}^{w \times h \times d}$ for I_n, where w, h, and d are the width, the height, and the number of channels, respectively. Similarly, feature maps $F_m \in \mathbb{R}^{w \times h \times d}$ are yielded for I_m. Our model is then divided into two streams. One stream performs correlation between F_n and F_m, and yields a 4D correlation tensor $T_{nm} \in \mathbb{R}^{w \times h \times w \times h}$. Each element $T_{nm}(i, j, s, t) = T_{nm}(\mathbf{p}, \mathbf{q})$ records the normalized inner product between the feature vectors stored at two spatial locations, i.e., $\mathbf{p} = [i, j]$ in F_n and $\mathbf{q} = [s, t]$ in F_m. The other stream employs a 1×1 convolutional layer to estimate the co-saliency map $\tilde{S}_k \in \mathbb{R}^{w \times h}$ of I_k and adopts deconvolution layers to generate a high-resolution co-saliency map $S_k \in \mathbb{R}^{W \times H}$, for $k \in \{n, m\}$. We design three loss functions, including the co-peak loss ℓ_t, the affinity loss ℓ_a, and the saliency loss ℓ_s, to derive the

network, leading to the following object function

$$\begin{aligned}\mathcal{L}(\mathbf{w}) = \lambda_t \sum_{n=1}^{N} \sum_{m \neq n} \ell_t(I_n, I_m; \mathbf{w}) \\ + \lambda_a \sum_{n=1}^{N} \sum_{m \neq n} \ell_a(I_n, I_m; \mathbf{w}) + \sum_{n=1}^{N} \ell_s(I_n; \mathbf{w}),\end{aligned} \tag{5.1}$$

where $\mathbf{w}$ is the set of learnable parameters of the network. Nonnegative weights λ_t and λ_a control the relative importance among the three losses. They are fixed to 0.5 and 0.1 in this work, respectively. The co-peak loss ℓ_t stimulates co-peak detection. The affinity loss ℓ_a refers to the co-saliency maps and enables discriminative feature learning. The saliency loss ℓ_s working with the other two losses carries out co-saliency detection and hence facilitates instance co-segmentation. The three losses are elaborated in the following.

Co-peak loss ℓ_t. This loss aims to stimulate co-peak detection. A co-peak consists of to two points, one in each of I_n and I_m. Since a co-peak covered by a pair of instances of the same object category is desired, the two points of the co-peak must be inside the object and similar to each other. Therefore, both *intra-image saliency* and *inter-image correlation* are taken into account in this loss.

As shown in Figure 5.1, our two-stream network produces the intra-image saliency maps $\tilde{S}_n$ and $\tilde{S}_m$ in one stream and inter-image correlation map T_{nm} in the other stream. To jointly consider the two types of information, a saliency-guided correlation tensor $T^s_{nm} \in \mathbb{R}^{w \times h \times w \times h}$ is constructed with its elements defined below

$$T^s_{nm}(\mathbf{p}, \mathbf{q}) = \tilde{S}_n(\mathbf{p}) \tilde{S}_m(\mathbf{q}) T_{nm}(\mathbf{p}, \mathbf{q}), \tag{5.2}$$

where $\mathbf{p} \in \mathcal{P}$, $\mathbf{q} \in \mathcal{P}$, and $\mathcal{P}$ is the set of all spatial coordinates of the feature maps. In Eq. (5.2), $\tilde{S}_n(\mathbf{p})$ is the saliency value of $\tilde{S}_n$ at point $\mathbf{p}$, and $\tilde{S}_m(\mathbf{q})$ is similarly defined.

To have more reliable keypoints to reveal object instances, we define a co-peak as a local maximum in T^s_{nm} within a 4D local window of size $3 \times 3 \times 3 \times 3$. Suppose that

$(\mathbf{p}, \mathbf{q})$ is a peak in T^s_{nm}. Both point $\mathbf{p}$ in F_n and point $\mathbf{q}$ in F_m are salient, and they are the most similar to each other in a local region. The former property implies that the two points probably reside in two salient object instances. The latter one reveals that the two instances are likely of the same class, since they have similar parts. Based on above discussion, the co-peak loss used to stimulate reliable co-peaks is defined by

$$\ell_t(I_n, I_m) = -log\left(\frac{1}{|\mathcal{M}_{nm}|}\sum_{(\mathbf{p},\mathbf{q})\in\mathcal{M}_{nm}} T^s_{nm}(\mathbf{p}, \mathbf{q})\right), \tag{5.3}$$

where $\mathcal{M}_{nm}$ is the set of co-peaks.

Affinity loss ℓ_a. The co-peak loss refers to the feature maps of the images, so discriminative features that can separate instances from background are preferable. Besides, the co-peak loss is applied to the locations of co-peaks, and features on other locations are ignored. The affinity loss is introduced to address the two issues. It aims to derive the features with which pixels in the salient regions are similar to each other while being distinct from those in the background. For a pair of images I_n and I_m, a loss $\tilde{\ell}_a(I_n, I_m)$ is defined by

$$\begin{aligned}\tilde{\ell}_a(I_n, I_m) = &\sum_{\mathbf{p}\in\mathcal{P}}\sum_{\mathbf{q}\in\mathcal{P}} \tilde{S}_n(\mathbf{p})\tilde{S}_m(\mathbf{q})\left(1 - T_{nm}(\mathbf{p}, \mathbf{q})\right) \\ &+ \alpha(\tilde{S}_n(\mathbf{p}) - \tilde{S}_m(\mathbf{q}))^2 T_{nm}(\mathbf{p}, \mathbf{q}),\end{aligned} \tag{5.4}$$

where constant α is empirically set to 4. In Eq. (5.4), the first term penalizes the case of low similarity between two salient pixels, while the second term prevents high similarity between a salient pixel and a non-salient pixel. The proposed affinity loss generalizes $\tilde{\ell}_a$ in Eq. (5.4) to consider both inter-image and intra-image affinities and is defined by

$$\ell_a(I_n, I_m) = \tilde{\ell}_a(I_n, I_m) + \tilde{\ell}_a(I_n, I_n) + \tilde{\ell}_a(I_m, I_m). \tag{5.5}$$

Saliency loss ℓ_s. This term aims to identify the salient regions and can guide the training of our model. Following the studies of object co-segmentation [65, 66, 124, 135], we utilize an off-the-shelf method for saliency detection. The resultant saliency maps can

serve as the object prior. In this work, we adopt the unsupervised method, SVFSal [162], which produces the saliency map $\hat{S}_n$ for image I_n. Note that the resolutions of $\hat{S}_n$ and I_n are the same. Thus, the deconvolutional layers are employed to increase the resolution. Following [60], the saliency loss ℓ_s applied to image I_n is defined by

$$\ell_s(I_n) = \sum_{\mathbf{p} \in I_n} \rho_n(\mathbf{p}) \| S_n(\mathbf{p}) - \hat{S}_n(\mathbf{p}) \|_2^2, \tag{5.6}$$

where $\mathbf{p}$ indexes the pixels of I_n, $\rho_n(\mathbf{p})$ is a weight representing the importance of pixel $\mathbf{p}$, and S_n is the predicted saliency map for I_n by our model. The weight $\rho_n(\mathbf{p})$ deals with the imbalance between the salient and non-salient areas. It is set to $1 - \varepsilon$ if pixel $\mathbf{p}$ resides in the salient region, and ε otherwise, where ε is the ratio of the salient area to the whole image. The mean value of $\hat{S}_n$ is used as the threshold to divide $\hat{S}_n$ into the salient and non-salient regions. In this way, the salient and non-salient regions contribute equally in Eq. (5.6). As shown in Figure 5.1, except for the deconvolutional layers, our model used to produce maps $\{S_n\}$ is derived by the three losses jointly. Thus, $\{S_n\}$ derived with both intra- and inter-image cues are called co-saliency maps. This prior term is helpful as it compensates for the lack of supervisory signals in instance co-segmentation.

5.2.3 Instance mask segmentation

After optimizing Eq. (5.1), we simply use the detected peaks on the estimated co-saliency maps as the final co-peaks, because detecting the co-peaks on all possible image pairs is complicated. Thus, the peaks $\{p_n^i\}_{i=1}^M$ of each image I_n are collected, where M is the number of the peaks. We adopt the method called *peak back-propagation* [171] to infer an instance-aware heat map O_n^i for each peak p_n^i. The map O_n^i is supposed to highlight the instance covering p_n^i. An example is given in Figure 5.1.

For instance mask generation, we utilize an unsupervised method, called *multi-scale combinatorial grouping* (MCG) [115], to produce a set of instance proposals for image I_n. With the heat maps $\{O_n^i\}_{i=1}^M$ and the co-saliency map S_n, we extend the proposal ranking function in [171] by further taking the co-saliency cues into account, and select the top-

ranked proposal as the mask for each detected peak. Specifically, given the maps O_n^i and S_n, the ranking function R applied to an instance proposal P is defined by

$$R(P) = \beta(O_n^i * S_n) * P + (O_n^i * S_n) * \hat{P} - \gamma(1 - S_n) * P, \tag{5.7}$$

where $\hat{P}$ is the contour of the proposal P and operator $*$ is the Frobenius inner product between two matrices. The coefficients β and γ are set to 0.8 and 10^{-5}, respectively. In Eq. (5.7), three terms, i.e., the instance-aware, contour-preserving, and object-irrelevant terms, are included. The instance-aware term prefers the proposals that cover the regions with high responses in O_n^i and high saliency in S_n. The contour-preserving term focuses on the fine-detailed boundary information. The background map, $1 - S_n$, is used in the object-irrelevant term to suppress background regions. Compared with the ranking function in [171], ours further exploits the properties of instance co-segmentation, i.e., the high co-saliency values in object instances, and can select more accurate proposals. Following a standard protocol of instance segmentation, we perform *non-maximum suppression* (NMS) to remove the redundancies.

5.2.4 Implementation details

We implement the proposed method using *MatConvNet* [141]. VGG-16 [131] is adopted as the feature extractor g. It is pre-trained on the ImageNet [29] dataset, and is updated during optimizing Eq. (5.1). The same network architecture is used in all experiments. Note that the objective in Eq. (5.1) involves all image pairs. Direct optimization is not feasible due to the limited memory size. Thereby, we adopt the *piecewise training* scheme [133]. Namely, only a subset of images is considered in each epoch, and the subset size is set to 6 in this work. The learning rate, weight decay, and momentum are set to 10^{-6}, 0.0005, and 0.9, respectively. The optimization procedure stops after 40 epochs. We choose ADAM [78] as the optimization solver. All images are resized to the resolution 448×448 in advance. We resize the instance co-segmentation results back to the original image resolution for performance evaluation.

dataset	(a)	(b)	(c)	(d)	(e)
COCO-VOC	12	1281	3151	106.8	2.5
COCO-NONVOC	32	3130	8303	91.8	2.7
VOC12	18	891	2214	178.2	2.5
SOC	5	522	835	29.0	1.6

Table 5.1: Some statistics of the four collected datasets, including (a) the number of classes, (b) the number of images, (c) the number of instances, (d) the average number of images per class, and (e) the average number of instances per image.

method	year	trained	COCO-VOC		COCO-NONVOC		VOC12		SOC	
			$mAP^r_{0.25}$	$mAP^r_{0.5}$	$mAP^r_{0.25}$	$mAP^r_{0.5}$	$mAP^r_{0.25}$	$mAP^r_{0.5}$	$mAP^r_{0.25}$	$mAP^r_{0.5}$
CLRW [134]	CVPR 2014	×	33.3	13.7	24.6	10.7	29.2	10.5	34.9	15.6
UODL [22]	CVPR 2015	×	9.6	2.2	8.5	1.8	9.4	2.0	11.0	2.7
DDT [149]	IJCAI 2017	×	31.4	10.1	25.7	9.7	30.7	8.8	43.0	25.7
DDT+ [150]	arXiv 2017	×	31.7	10.6	26.0	10.1	33.6	9.4	39.6	22.4
DFF [26]	ECCV 2018	×	30.8	11.6	22.6	7.3	27.7	13.7	42.3	17.0
NLDF [108]	CVPR 2017	√	39.1	18.2	23.9	8.5	34.3	12.7	49.5	21.6
C2S-Net [94]	ECCV 2018	√	39.6	13.4	25.1	7.6	30.1	10.7	37.0	12.5
PRM [171]	CVPR 2018	√	44.9	14.6	-	-	45.3	14.8	-	-
Ours	-	×	52.6	21.1	35.3	12.3	45.6	16.7	54.2	26.0

Table 5.2: Performance of instance co-segmentation on the four collected datasets. The numbers in red and green show the best and the second best results, respectively. The column "trained" indicates whether additional training data are used.

5.3 Experimental results

In this section, our method for instance co-segmentation and its variant for co-localization are evaluated. First, the adopted datasets and evaluation metrics are described. Then, the competing methods are introduced. Finally, the comparison results are reported and analyzed.

5.3.1 Dataset collection

As instance co-segmentation is a new task, no public benchmarks exist. Therefore, we establish four datasets with pixel-wise instance annotations by collecting images from three public benchmarks, including the MS COCO [98], PASCAL VOC 2012 [30, 41], and SOC [32] datasets. The following pre-processing is applied to each dataset. First, we remove the images where objects of more than one category are present. Second, we discard the categories that contain less than 10 images. The details of collecting images

from each dataset are described below.

MS COCO dataset. We collect images from the training and validation sets of the MS COCO 2017 object detection task. As MS CCCO is a large-scale dataset, we further remove the images that do not contain at least two instances. Total 44 categories remain. Some competing methods are pre-trained on PASCAL VOC 2012 dataset. For the ease of comparison, we divide the 44 categories into two disjoint sets, *COCO-VOC* and *COCO-NONVOC*. The former contains 12 categories covered by the PASCAL VOC 2012 dataset, while the latter contains the rest.

PASCAL VOC 2012 dataset. Because few pixel-wise instance annotations are available in the PASCAL VOC 2012 dataset, we adopt the augmented VOC12 dataset [41], which has 18 object categories after dataset preprocessing.

SOC dataset. SOC [32] is a newly collected dataset for saliency detection. It provides image-level labels and instance-aware annotations. After preprocessing, only five object categories remain because many images contain object instances of multiple categories and some categories have less than 10 images.

The statistics and the abbreviations of the four collected datasets are given in Table 5.1. Note that our method can work on images containing one or multiple instances of the common object category. The SOC dataset helps test this issue. As shown in Table 5.1, the average number of instances in SOC is 1.6, less than 2. It shows that there exist many images in this dataset with only one object instance. Please refer to the supplementary material for more details and some image samples of the four collected datasets.

5.3.2 Evaluation metrics

For instance co-segmentation, mean average precision (mAP) [42] is adopted as the performance measure. Following [171], we report mAP using the IoU thresholds at 0.25 and 0.5, denoted as $\text{mAP}^r_{0.25}$ and $\text{mAP}^r_{0.5}$, respectively.

For object co-localization, the performance measure CorLoc [22, 26, 134, 149, 150] is used as the evaluation metric. The measure CorLoc is designed for evaluating the results

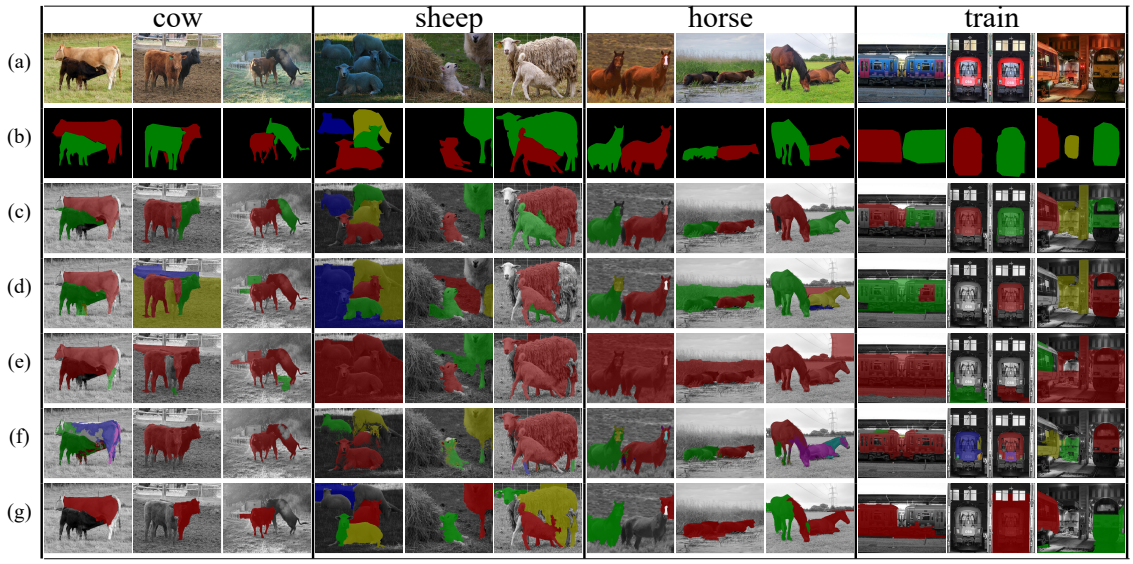


Figure 5.2: Results of instance co-segmentation on four object categories, i.e., *cow*, *sheep*, *horse*, and *train*, of the COCO-VOC dataset. (a) Input images. (b) Ground truth. (c) ∼ (g) Results with instance-specific coloring generated by different methods including (c) our method, (d) CLRW [134], (e) DFF [26], (f) NLDF [108], and (g) PRM [171], respectively.

in the form of object bounding boxes. For comparing with methods whose output is object or instance segments, we extend CorLoc to CorLoc^r to evaluate the results in the form of object segments.

5.3.3 Competing methods

As instance co-segmentation is a new task, there are no existing methods for performance comparison. We adopt two strategies for comparing our method with existing ones. First, we consider competing methods of three categories, including *object co-localization*, *class-agnostic saliency segmentation*, and *weakly supervised instance segmentation*. For methods of the three categories, we convert their predictions into the results in the form of instance co-segmentation, namely one segment mask for each detected instance. In this way, our method can be compared with these methods on the task of instance co-segmentation.

Second, we compare our method with methods of all the aforementioned three categories on the task of object co-localization. To this end, we need to convert the output of each compared method into the results in the form of object co-localization, namely the

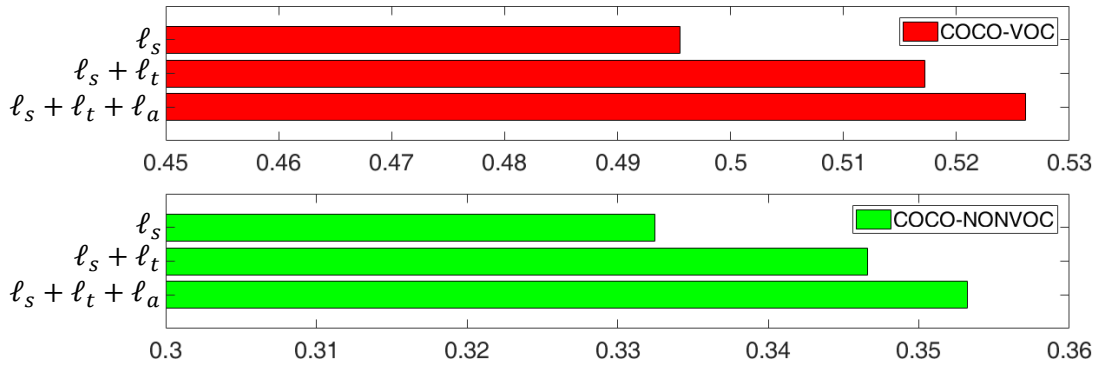


Figure 5.3: Performance in $\text{mAP}^r_{0.25}$ with different loss function combinations on the COCO-VOC and COCO-NONVOC datasets.

object bounding box with the highest confidence in each image.

In the two strategies of method comparison, two types of prediction conversion are required, including converting a bounding box to an instance segment and its inverse direction. Unless further specified, we adopt the following way to convert a bounding box prediction to an instance segment. Given a bounding box in an image, we apply MCG [115] to that image to generate a set of instance proposals, and retrieve the proposal with the highest IoU with the bounding box to represent it. On the other hand, it is easy to concert a given instance segment to a bounding box. We simply use the bounding box of that instance segment to represent it. In the following, the selected competing methods from each of the three categories are specified.

Object co-localization. We choose the state-of-the-art methods of this category for comparison, including CLRW [134], UODL [22], DDT [149], DDT+ [150], and DFF [26]. The first two methods, CLRW and UODL, output all bounding boxes with their scores, but cannot determine the number of instances in each image. Thus, we pick the top-scored bounding boxes as many as the instances detected by our method, and similarly apply NMS to remove redundancies. The last three methods, DDT, DDT+, and DFF, first produce the heat maps to highlight objects, then convert the heat maps into the binary masks by using their proposed mechanisms, and finally take the bounding boxes of the connected components on the binary masks.

	COCO-VOC		COCO-NONVOC	
	$mAP^r_{0.25}$	$mAP^r_{0.5}$	$mAP^r_{0.25}$	$mAP^r_{0.5}$
w/o co-saliency map	33.5	12.4	25.3	8.3
w co-saliency map	52.6	21.1	35.3	12.3

Table 5.3: Performance of our method working with the proposal ranking function without or with the co-saliency information on the COCO-VOC and COCO-NONVOC datasets.

Class-agnostic instance segmentation (CAIS). We select two powerful methods, NLDF [108] and C2S-Net [94], of this category as the competing methods. The algorithm proposed in [88] is used to convert the saliency contours generated by NLDF and C2S-Net into the results in the form of instance co-segmentation.

Weakly supervised instance segmentation (WSIS). The WSIS method, PRM [171], is trained on the PASCAL VOC 2012 dataset, and it cannot be applied to the images whose categories are not covered by the PASCAL VOC 2012 dataset. Therefore, PRM is compared with our method only on the COCO-VOC and VOC12 datasets.

5.3.4 Instance co-segmentation

For the ease of performance analysis, we divide the evaluated methods into two groups, i.e., trained and non-trained. The group trained includes NLDF [108], C2S-Net [94] and PRM [171]. Methods of this group require additional training data other than the input to instance co-segmentation. The other group non-trained contains our method and the rest of the competing methods. Methods of group non-trained have access to only the input to instance co-segmentation.

Our method and all competing methods are evaluated on the four collected datasets. Their performance is reported in Table 5.2. The proposed method outperforms the competing methods of group non-trained by large margins even though all of them access the same data. We attribute the performance gain yielded by our method to feature learning enabled CNNs. The competing methods of group non-trained adopt pre-defined features, and cannot well deal with complex and diverse intra-class variations and background clutters. On the contrary, our method leverages CNNs to carry out feature learning and in-

stance co-segmentation simultaneously, leading to much better performance. Although the methods of group trained have access to additional training data, ours still reaches more favorable results. The main reason is that our method explores co-occurrent patterns via co-peak detection when images for instance co-segmentation are available, while the methods of group trained fix their models after training on additional data and cannot adapt themselves to newly given images for instance co-segmentation.

To gain the insight into the quantitative results, Figure 5.2 visualizes the qualitative results generated by our method, CLRW [134], DFF [26], NLDF [108], and PRM [171]. The major difficulties of instance segmentation lie in instance mutual occlusions, intra-class variations, and cluttered scene. As shown in Figure 5.2(c), our method still works well when instance mutual occlusions occur on categories *cow*, *sheep*, and *horse* and large intra-class variations and cluttered scene are present on category *train*. In Figure 5.2(d), CLRW yields some false alarms in the background while has false negatives on category *train*. In Figure 5.2(e), DFF cannot well address instance mutual occlusions due to computing connected components for instance identification. In Figure 5.2(f) and Figure 5.2(g), NLDF and CRP perform favorably against other competing methods, but still suffer from over-segmentation and misses, respectively.

Ablation studies. We analyze the proposed objective consisting of three loss functions in Eq. (5.1) on the COCO-VOC and COCO-NONVOC datasets, and report the results in Figure 5.3. Except loss ℓ_s, the other two losses, ℓ_t and ℓ_a, are added one by one. When ℓ_t is included, the performance gains are significant on both datasets. It implies that ℓ_t for reliable co-peak search is important in our method. Once ℓ_a is added, the performance is moderately enhanced, which means that discriminative feature learning is helpful for instance co-segmentation. In addition to the objective, the effect of referring to co-saliency maps in proposal ranking is analyzed in Table 5.3. The results clearly point out that information from co-saliency detection is crucial to proposal ranking. It is not surprised. Since co-peaks identify the keypoints within instances, we still need the evidence from co-saliency maps to reveal the corresponding instances.

method	year	trained	COCO-VOC	COCO-NONVOC	VOC12	SOC
CLRW [134]	CVPR 2014	×	33.4	31.6	29.9	30.9
UODL [22]	CVPR 2015	×	12.3	12.7	9.5	10.3
DDT [149]	IJCAI 2017	×	30.0	27.4	25.0	16.7
DDT+ [150]	PR 2019	×	29.5	25.8	23.7	18.4
DFF [26]	ECCV 2018	×	32.3	30.5	28.7	22.9
NLDF [108]	CVPR 2017	√	51.2	31.0	39.2	42.0
C2S-Net [94]	ECCV 2018	√	39.0	28.4	31.1	32.9
PRM [171]	CVPR 2018	√	18.1	-	23.3	-
Ours	-	×	49.6	34.3	39.2	43.1

Table 5.4: Performance of object co-localization on the four datasets. The numbers in red and green indicate the best and the second best results, respectively. The column "trained" indicates whether additional training data are used.

5.3.5 Object co-localization

We evaluate our method and the competing methods for object co-localization in the four datasets we collected. For our method, we pick the top-ranked proposal in each image when evaluating the performance in CorLoc^r. Table 5.4 reports the performance of all the compared methods. Our method achieves the comparable or even better performance, even though it is not originally designed for object co-localization. Seven examples of object co-localization by our method are shown in Figure 5.4, where accurate instance masks and the corresponding bounding boxes are discovered by our method.

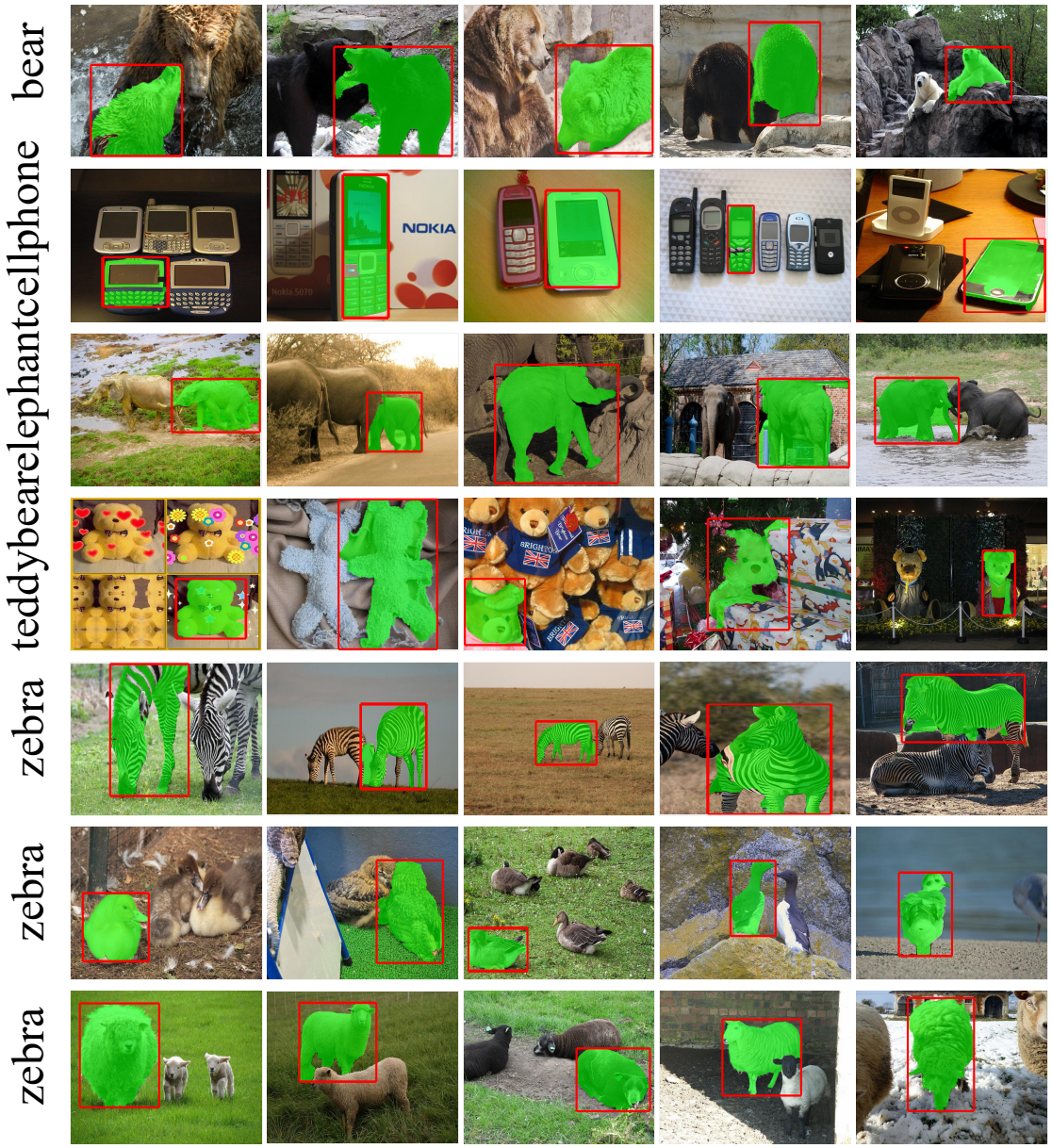


Figure 5.4: Seven examples, one in each row, of the co-localization results by our method on the COCO-NONVOC dataset.

Chapter 6

Conclusion and Future Work

We conclude the thesis and discuss future work in this chapter.

6.1 Conclusion

In this thesis, we have developed four CNN-based approaches for the visual attention-getting object discovery without the pixel-wise annotations. We address four tasks related to the visual attention-getting object discovery, including the top-down saliency detection, object co-segmentation, object co-saliency detection, and . Current state-of-the-art approaches in tasks require the pixel-wise annotations as the training data to learn their proposed models. However, pixel-wise annotations are usually manually drawn or delineated by tools with intensive user interaction, and the massive annotation cost for collecting the training data reduces the flexibility and generalization of the visual attention-getting object discovery. In this thesis, our goal is to propose a series of methods to train a CNN-based model with the minimum annotation effort.

First, 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.

Second, we have presented a CNN-based approach for object co-segmentation task without pixel-level annotation pre-training. The proposed CNN architecture is composed of two CNN modules, *a feature extractor* and *a co-attention map generator*, along with two unsupervised losses, *a co-attention loss* and *a mask loss*. During the optimization process, the similarity of estimated objects and background is calculated in the co-attention loss, and the information can be propagated to guide the optimization of the generator. Thus, the co-attention loss can enable the generator to produce maps correctly localizing the common objects. The mask loss further regularizes the optimization. The mask loss can regularize the generator to remove false negatives and positives on objects and background, respectively. The experimental results on three challenging benchmarks are promising, and the proposed method outperforms the existing state-of-the-art unsupervised and supervised approaches.

Third, we have presented a method built on CNNs for co-saliency detection. Our approach decomposes the problem into two sub-tasks, *single-image saliency detection* and *cross-image co-occurrence region discovery*, by modeling the corresponding novel losses: *single-image saliency (SIS) loss* and *co-occurrence (COOC) loss*. The graphical model is adopted to integrate these two losses with unary and pairwise terms corresponding to the SIS and COOC losses, respectively. By optimizing the energy function associated with the graph, two networks are learned jointly. The quality of co-saliency maps is further improved by self-paced learning and postprocessing by DenseCRFs. Experiments on three challenging benchmarks show that the proposed method outperforms the state-of-the-art unsupervised methods.

Finally, we have presented an interesting and challenging task called instance co-segmentation, and propose a CNN-based method to effectively solve it without using additional training data. We decompose this task into two sub-tasks, including co-peak search and instance mask segmentation. In the former sub-task, we design three novel losses, co-peak, affinity, and saliency losses, for joint co-peak and co-saliency map detection. In the latter sub-task, we develop an effective proposal ranking algorithm, and can retrieve high-quality proposals to accomplish instance co-segmentation. Our method for instance co-segmentation and its variant for object co-localization are extensively evaluated on the four collected datasets. Both quantitative and qualitative results show that our method and its variant perform favorably against the state-of-the-arts.

6.2 Future Work

There exist two potential research directions worth exploration, video saliency detection and instance-level semantic matching. In video saliency detection, current state-of-the-art methods are learning-based, so a lot of pixel-level annotations are needed to learn their proposed CNN architectures. The proposed co-attention losses in Chapter 3 and Chapter 4, can't straightforward be applied in this task because the background across the different frames is very similar. Based on the mutual reference across the different images, the proposed co-attention loss can't identify the salient objects in a single video. However, the temporal consistency is useful to identify the salient object in a video, and it should be integrated into the proposed co-attention loss.

In the second direction, the semantic matching [55, 62, 17, 19, 120] aims to densely identify pixel correspondences across images. Because of CNNs, the significant progress has been made on this task. However, current state-of-the-art methods, such as [19, 120], also require the true pixel correspondences to learn their proposed model. In addition, the instance-level semantic matching has not been explored. Our proposed co-peak loss proposed in Chapter 5 could be integrated [19, 120] to search the matching between different instances with an online fashion.

Bibliography

[1] Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurélien Lucchi, Pascal Fua, and Sabine Süsstrunk. SLIC superpixels compared to state-of-the-art superpixel methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(11):2274–2282, 2012.

[2] David Aldavert, Arnau Ramisa, Ramón López de Mántaras, and Ricardo Toledo. Fast and robust object segmentation with the integral linear classifier. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1046–1053, 2010.

[3] Min Bai and Raquel Urtasun. Deep watershed transform for instance segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2858–2866, 2017.

[4] Dhruv Batra, Adarsh Kowdle, Devi Parikh, Jiebo Luo, and Tsuhan Chen. iCoseg: Interactive co-segmentation with intelligent scribble guidance. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3169–3176, 2010.

[5] Archith John Bency, Heesung Kwon, Hyungtae Lee, S. Karthikeyan, and B. S. Manjunath. Weakly supervised localization using deep feature maps. In *Proceedings of the European Conference on Computer Vision*, pages 714–731, 2016.

[6] Hakan Bilen and Andrea Vedaldi. Weakly supervised deep detection networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2846–2854, 2016.

[7] Ali Borji, Ming-Ming Cheng, Huaizu Jiang, and Jia Li. Salient object detection: A survey. *arXiv*, 2014.

[8] Ali Borji, Ming-Ming Cheng, Huaizu Jiang, and Jia Li. Salient object detection: A benchmark. *IEEE Transactions on Image Processing*, 24(12):5706–5722, 2015.

[9] Chunshui Cao, Xianming Liu, Yi Yang, Yinan Yu, Jiang Wang, Zilei Wang, Yongzhen Huang, Liang Wang, Chang Huang, Wei Xu, Deva Ramanan, and Thomas S. Huang. Look and think twice: Capturing top-down visual attention with feedback convolutional neural networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2956–2964, 2015.

[10] Xiaochun Cao, Zhiqiang Tao, Bao Zhang, Huazhu Fu, and Wei Feng. Self-adaptively weighted co-saliency detection via rank constraint. *IEEE Transactions on Image Processing*, 23(9):4175–4186, 2014.

[11] Siddhartha Chandra, Nicolas Usunier, and Iasonas Kokkinos. Dense and low-rank gaussian CRFs using deep embeddings. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5113–5122, 2017.

[12] Feng-Ju Chang, Yen-Yu Lin, and Kuang-Jui Hsu. Multiple structured-instance learning for semantic segmentation with uncertain training data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 360–367, 2014.

[13] Haw-Shiuan Chang and Yu-Chiang Frank Wang. Optimizing the decomposition for multiple foreground cosegmentation. *Computer Vision and Image Understanding*, 141:18–27, 2015.

[14] Kai-Yueh Chang, Tyng-Luh Liu, Hwann-Tzong Chen, and Shang-Hong Lai. Fusing generic objectness and visual saliency for salient object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 914–921, 2011.

[15] Kai-Yueh Chang, Tyng-Luh Liu, and Shang-Hong Lai. From co-saliency to co-segmentation: An efficient and fully unsupervised energy minimization model. In

Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 2129–2136, 2011.

[16] Arslan Chaudhry, Puneet Kumar Dokania, and Philip H. S. Torr. Discovering class-specific pixels for weakly-supervised semantic segmentation. In *Proceedings of the British Machine Vision Conference*, 2017.

[17] Hsin-Yi Chen, Yen-Yu Lin, and Bing-Yu Chen. Co-segmentation guided Hough transform for robust feature matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(12):2388–2401, 2015.

[18] Xinlei Chen, Abhinav Shrivastava, and Abhinav Gupta. Enriching visual knowledge bases via object discovery and segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2035–2042, 2014.

[19] Yun-Chun Chen, Po-Hsiang Huang, Li-Yu Yu, Jia-Bin Huang, Ming-Hsuan Yang, and Yen-Yu Lin. Deep semantic matching with foreground detection and cycle-consistency. In *Proceedings of the Asian Conference on Computer Vision*, pages 347–362, 2018.

[20] Jie-Zhi Cheng, Feng-Ju Chang, Kuang-Jui Hsu, and Yen-Yu Lin. Knowledge leverage from contours to bounding boxes: A concise approach to annotation. In *Proceedings of the Asian Conference on Computer Vision*, pages 730–744, 2012.

[21] Ming-Ming Cheng, Niloy J. Mitra, Xiaolei Huang, Philip H. S. Torr, and Shi-Min Hu. Global contrast based salient region detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(3):569–582, 2015.

[22] Minsu Cho, Suha Kwak, Cordelia Schmid, and Jean Ponce. Unsupervised object discovery and localization in the wild: Part-based matching with bottom-up region proposals. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1201–1210, 2015.

[23] Hisham Cholakkal, Jubin Johnson, and Deepu Rajan. Backtracking ScSPM image classifier for weakly supervised top-down saliency. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5278–5287, 2016.

[24] Hisham Cholakkal, Jubin Johnson, and Deepu Rajan. Backtracking spatial pyramid pooling-based image classifier for weakly supervised top-down salient object detection. *IEEE Transactions on Image Processing*, 27(12):6064–6078, 2018.

[25] Hisham Cholakkal, Deepu Rajan, and Jubin Johnson. Top-down saliency with locality-constrained contextual sparse coding. In *Proceedings of the British Machine Vision Conference*, pages 159.1–159.12, 2015.

[26] Edo Collins, Radhakrishna Achanta, and Sabine Süsstrunk. Deep feature factorization for concept discovery. In *Proceedings of the European Conference on Computer Vision*, pages 352–368, 2018.

[27] Jifeng Dai, Kaiming He, and Jian Sun. Instance-aware semantic segmentation via multi-task network cascades. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3150–3158, 2016.

[28] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 886–893, 2005.

[29] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. ImageNet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[30] Mark Everingham, Luc J. Van Gool, Christopher K. I. Williams, John M. Winn, and Andrew Zisserman. The pascal visual object classes (VOC) challenge. *International Journal of Computer Vision*, 88(2):303–338, 2010.

[31] Alon Faktor and Michal Irani. Co-segmentation by composition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1297–1304, 2013.

[32] Deng-Ping Fan, Ming-Ming Cheng, Jiangjiang Liu, Shanghua Gao, Qibin Hou, and Ali Borji. Salient objects in clutter: Bringing salient object detection to the foreground. In *Proceedings of the European Conference on Computer Vision*, pages 196–212, 2018.

[33] Deng-Ping Fan, Ming-Ming Cheng, Yun Liu, Tao Li, and Ali Borji. Structure-Measure: A new way to evaluate foreground maps. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4558–4567, 2017.

[34] Ruochen Fan, Qibin Hou, Ming-Ming Cheng, Tai-Jiang Mu, and Shi-Min Hu. S^4net: Single stage salient-instance segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[35] Huazhu Fu, Xiaochun Cao, and Zhuowen Tu. Cluster-based co-saliency detection. *IEEE Transactions on Image Processing*, 22(10):3766–3778, 2013.

[36] Huazhu Fu, Dong Xu, Bao Zhang, Stephen Lin, and Rabab K. Ward. Object-based multiple foreground video co-segmentation via multi-state selection graph. *IEEE Transactions on Image Processing*, 24(11):3415–3424, 2015.

[37] Brian Fulkerson, Andrea Vedaldi, and Stefano Soatto. Class segmentation and object localization with superpixel neighborhoods. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 670–677, 2009.

[38] Clément Godard, Oisin Mac Aodha, and Gabriel J. Brostow. Unsupervised monocular depth estimation with left-right consistency. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6602–6611, 2017.

[39] Junwei Han, Gong Cheng, Zhenpeng Li, and Dingwen Zhang. A unified metric learning-based framework for co-saliency detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 28(10):2473–2483, 2018.

[40] Junwei Han, Rong Quan, Dingwen Zhang, and Feiping Nie. Robust object co-segmentation using background prior. *IEEE Transactions on Image Processing*, 27(4):1639–1651, 2018.

[41] Bharath Hariharan, Pablo Arbelaez, Lubomir D. Bourdev, Subhransu Maji, and Jitendra Malik. Semantic contours from inverse detectors. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 991–998, 2011.

[42] Bharath Hariharan, Pablo Andrés Arbeláez, Ross B. Girshick, and Jitendra Malik. Simultaneous detection and segmentation. In *Proceedings of the European Conference on Computer Vision*, pages 297–312, 2014.

[43] Avik Hati, Subhasis Chaudhuri, and Rajbabu Velmurugan. Image co-segmentation using maximum common subgraph matching and region co-growing. In *Proceedings of the European Conference on Computer Vision*, pages 736–752, 2016.

[44] Zeeshan Hayder, Xuming He, and Mathieu Salzmann. Learning to co-generate object proposals with a deep structured network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2565–2573, 2016.

[45] Zeeshan Hayder, Xuming He, and Mathieu Salzmann. Boundary-aware instance segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 587–595, 2017.

[46] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. Mask R-CNN. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2980–2988, 2017.

[47] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[48] Shengfeng He and Rynson W. H. Lau. Exemplar-driven top-down saliency detection via deep association. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5723–5732, 2016.

[49] Shengfeng He, Rynson W. H. Lau, Wenxi Liu, Zhe Huang, and Qingxiong Yang. Supercnn: A superpixelwise convolutional neural network for salient object detection. *International Journal of Computer Vision*, 115(3):330–344, 2015.

[50] Seunghoon Hong, Tackgeun You, Suha Kwak, and Bohyung Han. Online tracking by learning discriminative saliency map with convolutional neural network. In *Proceedings of the International Conference on Machine Learning*, pages 597–606, 2015.

[51] Qibin Hou, Ming-Ming Cheng, Xiaowei Hu, Ali Borji, Zhuowen Tu, and Philip H. S. Torr. Deeply supervised salient object detection with short connections. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5300–5309, 2017.

[52] Qibin Hou, Daniela Massiceti, Puneet Kumar Dokania, Yunchao Wei, Ming-Ming Cheng, and Philip H. S. Torr. Bottom-up top-down cues for weakly-supervised semantic segmentation. In *Proceedings of the International Conference on Energy Minimization Methods in Computer Vision and Pattern Recognition*, pages 263–277, 2017.

[53] Qinbin Hou, Puneet Kumar Dokania, Daniela Massiceti, Yunchao Wei, Ming-Ming Cheng, and Philip H. S. Torr. Mining pixels: Weakly supervised semantic segmentation using image labels. *arXiv*, 2016.

[54] Kuang-Jui Hsu, Yen-Yu Lin, and Yung-Yu Chuang. Augmented multiple instance regression for inferring object contours in bounding boxes. *IEEE Transactions on Image Processing*, 23(4):1722–1736, 2014.

[55] Kuang-Jui Hsu, Yen-Yu Lin, and Yung-Yu Chuang. Robust image alignment with multiple feature descriptors and matching-guided neighborhoods. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1921–1930, 2015.

[56] Kuang-Jui Hsu, Yen-Yu Lin, and Yung-Yu Chuang. Weakly supervised saliency detection with A category-driven map generator. In *Proceedings of the British Machine Vision Conference*, 2017.

[57] Kuang-Jui Hsu, Yen-Yu Lin, and Yung-Yu Chuang. Co-attention cnns for unsupervised object co-segmentation. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 748–756, 2018.

[58] Kuang-Jui Hsu, Yen-Yu Lin, and Yung-Yu Chuang. Deepco^3: Deep instance co-segmentation by co-peak search and co-saliency detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[59] Kuang-Jui Hsu, Yen-Yu Lin, and Yung-Yu Chuang. Weakly supervised salient object detection by learning a classifier-driven map generator. *IEEE Transactions on Image Processing*, 2019.

[60] Kuang-Jui Hsu, Chung-Chi Tsai, Yen-Yu Lin, Xiaoning Qian, and Yung-Yu Chuang. Unsupervised cnn-based co-saliency detection with graphical optimization. In *Proceedings of the European Conference on Computer Vision*, pages 502–518, 2018.

[61] Yuan-Ting Hu, Jia-Bin Huang, and Alexander G. Schwing. MaskRNN: Instance level video object segmentation. In *Proceedings of the Advances in Neural Information Processing Systems*, pages 324–333, 2017.

[62] Yuan-Ting Hu, Yen-Yu Lin, Hsin-Yi Chen, Kuang-Jui Hsu, and Bing-Yu Chen. Matching images with multiple descriptors: An unsupervised approach for locally adaptive descriptor selection. *IEEE Transactions on Image Processing*, 24(12):5995–6010, 2015.

[63] Chun-Rong Huang, Yun-Jung Chang, Zhi-Xiang Yang, and Yen-Yu Lin. Video saliency map detection by dominant camera motion removal. *IEEE Transactions on Circuits and Systems for Video Technology*, 24(8):1336–1349, 2014.

[64] Fang Huang, Jinqing Qi, Huchuan Lu, Lihe Zhang, and Xiang Ruan. Salient object detection via multiple instance learning. *IEEE Transactions on Image Processing*, 26(4):1911–1922, 2017.

[65] Koteswar Rao Jerripothula, Jianfei Cai, Jiangbo Lu, and Junsong Yuan. Object co-skeletonization with co-segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3881–3889, 2017.

[66] Koteswar Rao Jerripothula, Jianfei Cai, and Junsong Yuan. Image co-segmentation via saliency co-fusion. *IEEE Transactions on Multimedia*, 18(9):1896–1909, 2016.

[67] Yangqing Jia and Mei Han. Category-independent object-level saliency detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1761–1768, 2013.

[68] Bowen Jiang, Lihe Zhang, Huchuan Lu, Chuan Yang, and Ming-Hsuan Yang. Saliency detection via absorbing markov chain. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1665–1672, 2013.

[69] Huaizu Jiang, Jingdong Wang, Zejian Yuan, Tie Liu, and Nanning Zheng. Automatic salient object segmentation based on context and shape prior. In *Proceedings of the British Machine Vision Conference*, pages 110.1–110.12, 2011.

[70] Peng Jiang, Haibin Ling, Jingyi Yu, and Jingliang Peng. Salient region detection by UFO: Uniqueness, focusness and objectness. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1976–1983, 2013.

[71] Peng Jiang, Nuno Vasconcelos, and Jingliang Peng. Generic promotion of diffusion-based salient object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 217–225, 2015.

[72] Bin Jin, Maria V. Ortiz Segovia, and Sabine Süsstrunk. Webly supervised semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1705–1714, 2017.

[73] Armand Joulin, Francis R. Bach, and Jean Ponce. Discriminative clustering for image co-segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1943–1950, 2010.

[74] Armand Joulin, Francis R. Bach, and Jean Ponce. Multi-class cosegmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 542–549, 2012.

[75] Vadim Kantorov, Maxime Oquab, Minsu Cho, and Ivan Laptev. ContextLocNet: Context-aware deep network models for weakly supervised localization. In *Proceedings of the European Conference on Computer Vision*, pages 350–365, 2016.

[76] Gunhee Kim, Eric P. Xing, Fei-Fei Li, and Takeo Kanade. Distributed cosegmentation via submodular optimization on anisotropic diffusion. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 169–176, 2011.

[77] Jiwhan Kim, Dongyoon Han, Yu-Wing Tai, and Junmo Kim. Salient region detection via high-dimensional color transform. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 883–890, 2014.

[78] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations*, 2015.

[79] Patrick Knöbelreiter, Christian Reinbacher, Alexander Shekhovtsov, and Thomas Pock. End-to-end training of hybrid CNN-CRF models for stereo. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1456–1465, 2017.

[80] Aysun Kocak, Kemal Cizmeciler, Aykut Erdem, and Erkut Erdem. Top down saliency estimation via superpixel-based discriminative dictionaries. In *Proceedings of the British Machine Vision Conference*, 2014.

[81] Alexander Kolesnikov and Christoph H. Lampert. Seed, expand and constrain: Three principles for weakly-supervised image segmentation. In *Proceedings of the European Conference on Computer Vision*, pages 695–711, 2016.

[82] Shu Kong and Charless C. Fowlkes. Recurrent pixel embedding for instance grouping. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9018–9028, 2018.

[83] Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected CRFs with gaussian edge potentials. In *Proceedings of the Advances in Neural Information Processing Systems*, pages 109–117, 2011.

[84] Philipp Krähenbühl and Vladlen Koltun. Geodesic object proposals. In *Proceedings of the European Conference on Computer Vision*, pages 725–739, 2014.

[85] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. ImageNet classification with deep convolutional neural networks. In *Proceedings of the Advances in Neural Information Processing Systems*, pages 1106–1114, 2012.

[86] Chulwoo Lee, Won-Dong Jang, Jae-Young Sim, and Chang-Su Kim. Multiple random walkers and their application to image cosegmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3837–3845, 2015.

[87] Dong Li, Jia-Bin Huang, Yali Li, Shengjin Wang, and Ming-Hsuan Yang. Weakly supervised object localization with progressive domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3512–3520, 2016.

[88] Guanbin Li, Yuan Xie, Liang Lin, and Yizhou Yu. Instance-level salient object segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 247–256, 2017.

[89] Guanbin Li and Yizhou Yu. Deep contrast learning for salient object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 478–487, 2016.

[90] Lina Li, Zhi Liu, and Jian Zhang. Unsupervised image co-segmentation via guidance of simple images. *Neurocomputing*, 275:1650–1661, 2018.

[91] Shuang Li, Huchuan Lu, Zhe L. Lin, Xiaohui Shen, and Brian L. Price. Adaptive metric learning for saliency detection. *IEEE Transactions on Image Processing*, 24(11):3321–3331, 2015.

[92] Weihao Li, Omid Hosseini Jafari, and Carsten Rother. Deep object co-segmentation. In *Proceedings of the Asian Conference on Computer Vision*, pages 638–653, 2018.

[93] Xi Li, Liming Zhao, Lina Wei, Ming-Hsuan Yang, Fei Wu, Yueting Zhuang, Haibin Ling, and Jingdong Wang. Deepsaliency: Multi-task deep neural network model for salient object detection. *IEEE Transactions on Image Processing*, 25(8):3919–3930, 2016.

[94] Xin Li, Fan Yang, Hong Cheng, Wei Liu, and Dinggang Shen. Contour knowledge transfer for salient object detection. In *Proceedings of the European Conference on Computer Vision*, pages 370–385, 2018.

[95] Yi Li, Haozhi Qi, Jifeng Dai, Xiangyang Ji, and Yichen Wei. Fully convolutional instance-aware semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4438–4446, 2017.

[96] YiJun Li, Keren Fu, Zhi Liu, and Jie Yang. Efficient saliency-model-guided visual co-saliency detection. *IEEE Signal Processing Letters*, 22(5):588–592, 2015.

[97] Xiaodan Liang, Liang Lin, Yunchao Wei, Xiaohui Shen, Jianchao Yang, and Shuicheng Yan. Proposal-free network for instance-level object segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(12):2978–2991, 2018.

[98] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In *Proceedings of the European Conference on Computer Vision*, pages 740–755, 2014.

[99] Fayao Liu, Chunhua Shen, and Guosheng Lin. Deep convolutional neural fields for depth estimation from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5162–5170, 2015.

[100] Feng Liu and Michael Gleicher. Region enhanced scale-invariant saliency detection. In *Proceedings of the IEEE International Conference on Multimedia and Expo*, pages 1477–1480, 2006.

[101] Nian Liu and Junwei Han. DHSNet: Deep hierarchical saliency network for salient object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 678–686, 2016.

[102] Nian Liu, Junwei Han, and Ming-Hsuan Yang. PiCANet: Learning pixel-wise contextual attention for saliency detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3089–3098, 2018.

[103] Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia. Path aggregation network for instance segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8759–8768, 2018.

[104] Yiding Liu, Siyu Yang, Bin Li, Wengang Zhou, Jizheng Xu, Houqiang Li, and Yan Lu. Affinity derivation and graph merge for instance segmentation. In *Proceedings of the European Conference on Computer Vision*, pages 708–724, 2018.

[105] Zhi Liu, Wenbin Zou, Lina Li, Liquan Shen, and Olivier Le Meur. Co-saliency detection based on hierarchical segmentation. *IEEE Signal Processing Letters*, 36(1):88–92, 2014.

[106] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3431–3440, 2015.

[107] David G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[108] Zhiming Luo, Akshaya Kumar Mishra, Andrew Achkar, Justin A. Eichel, Shaozi Li, and Pierre-Marc Jodoin. Non-local deep features for salient object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6593–6601, 2017.

[109] Yu-Fei Ma and HongJiang Zhang. Contrast-based image attention analysis by using fuzzy growing. In *Proceedings of the Annual ACM Conference on Multimedia Conference*, pages 374–381, 2003.

[110] Marcin Marszalek and Cordelia Schmid. Accurate object recognition with shape masks. *International Journal of Computer Vision*, 97(2):191–209, 2012.

[111] Simon Meister, Junhwa Hur, and Stefan Roth. UnFlow: Unsupervised learning of optical flow with a bidirectional census loss. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7251–7259, 2018.

[112] David Novotný, Samuel Albanie, Diane Larlus, and Andrea Vedaldi. Semi-convolutional operators for instance segmentation. In *Proceedings of the European Conference on Computer Vision*, pages 89–105, 2018.

[113] Andreas Opelt, Axel Pinz, Michael Fussenegger, and Peter Auer. Generic object recognition with boosting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(3):416–431, 2006.

[114] Maxime Oquab, Léon Bottou, Ivan Laptev, and Josef Sivic. Is object localization for free? - Weakly-supervised learning with convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 685–694, 2015.

[115] Jordi Pont-Tuset, Pablo Arbelaez, Jonathan T. Barron, Ferran Marqués, and Jitendra Malik. Multiscale combinatorial grouping for image segmentation and object proposal generation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(1):128–140, 2017.

[116] Rong Quan, Junwei Han, Dingwen Zhang, and Feiping Nie. Object co-segmentation via graph optimized-flexible manifold ranking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 687–695, 2016.

[117] Scott E. Reed, Honglak Lee, Dragomir Anguelov, Christian Szegedy, Dumitru Erhan, and Andrew Rabinovich. Training deep neural networks on noisy labels with bootstrapping. In *Workshops of the International Conference on Learning Representations*, 2015.

[118] Zhe Ren, Junchi Yan, Bingbing Ni, Bin Liu, Xiaokang Yang, and Hongyuan Zha. Unsupervised deep learning for optical flow estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1495–1501, 2017.

[119] Zhixiang Ren, Shenghua Gao, Liang-Tien Chia, and Ivor Wai-Hung Tsang. Region-based saliency detection and its application in object recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 24(5):769–779, 2014.

[120] Ignacio Rocco, Relja Arandjelovic, and Josef Sivic. End-to-end weakly-supervised semantic alignment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6917–6925, 2018.

[121] Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. "GrabCut" - Interactive foreground extraction using iterated graph cuts. *ACM Transactions on Graphics*, 23(3):309–314, 2004.

[122] Carsten Rother, Thomas P. Minka, Andrew Blake, and Vladimir Kolmogorov. Cosegmentation of image pairs by histogram matching - incorporating a global constraint into mrfs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 993–1000, 2006.

[123] Anirban Roy and Sinisa Todorovic. Combining bottom-up, top-down, and smoothness cues for weakly supervised image segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7282–7291, 2017.

[124] Michael Rubinstein, Armand Joulin, Johannes Kopf, and Ce Liu. Unsupervised joint object discovery and segmentation in internet images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1939–1946, 2013.

[125] José C. Rubio, Joan Serrat, Antonio M. López, and Nikos Paragios. Unsupervised co-segmentation through region matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 749–756, 2012.

[126] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 618–626, 2017.

[127] Falong Shen, Rui Gan, Shuicheng Yan, and Gang Zeng. Semantic segmentation via structured patch prediction, context CRF and guidance CRF. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5178–5186, 2017.

[128] Wataru Shimoda and Keiji Yanai. Distinct class-specific saliency maps for weakly supervised semantic segmentation. In *Proceedings of the European Conference on Computer Vision*, pages 218–234, 2016.

[129] Jamie Shotton, John M. Winn, Carsten Rother, and Antonio Criminisi. Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context. *International Journal of Computer Vision*, 81(1):2–23, 2009.

[130] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *Workshops of the International Conference on Learning Representations*, 2014.

[131] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *Proceedings of the International Conference on Learning Representations*, 2015.

[132] Jian Sun and Jean Ponce. Learning dictionary of discriminative part detectors for image categorization and cosegmentation. *International Journal of Computer Vision*, 120(2):111–133, 2016.

[133] Charles A. Sutton and Andrew McCallum. Piecewise training for structured prediction. *Machine Learning*, 77(2-3):165–194, 2009.

[134] Kevin D. Tang, Armand Joulin, Li-Jia Li, and Fei-Fei Li. Co-localization in real-world images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1464–1471, 2014.

[135] Zhiqiang Tao, Hongfu Liu, Huazhu Fu, and Yun Fu. Image cosegmentation via saliency-guided constrained clustering with cosine similarity. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4285–4291, 2017.

[136] Chung-Chi Tsai, Weizhi Li, Kuang-Jui Hsu, Xiaoning Qian, and Yen-Yu Lin. Image co-saliency detection and co-segmentation via progressive joint optimization. *IEEE Transactions on Image Processing*, 28(1):56–71, 2019.

[137] Chung-Chi Tsai, Xiaoning Qian, and Yen-Yu Lin. Image co-saliency detection via locally adaptive saliency map fusion. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1897–1901, 2017.

[138] Chung-Chi Tsai, Xiaoning Qian, and Yen-Yu Lin. Segmentation guided local proposal fusion for co-saliency detection. In *Proceedings of the IEEE International Conference on Multimedia and Expo*, pages 523–528, 2017.

[139] Wei-Chih Tu, Shengfeng He, Qingxiong Yang, and Shao-Yi Chien. Real-time salient object detection with a minimum spanning tree. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2334–2342, 2016.

[140] Andrea Vedaldi and Brian Fulkerson. Vlfeat: an open and portable library of computer vision algorithms. In *Proceedings of the Annual ACM Conference on Multimedia Conference*, pages 1469–1472, 2010.

[141] Andrea Vedaldi and Karel Lenc. MatConvNet: Convolutional neural networks for MATLAB. In *Proceedings of the Annual ACM Conference on Multimedia Conference*, pages 689–692, 2015.

[142] Chuan Wang, Hua Zhang, Liang Yang, Xiaochun Cao, and Hongkai Xiong. Multiple semantic matching on augmented n-partite graph for object co-segmentation. *IEEE Transactions on Image Processing*, 26(12):5825–5839, 2017.

[143] Jingdong Wang, Huaizu Jiang, Zejian Yuan, Ming-Ming Cheng, Xiaowei Hu, and Nanning Zheng. Salient object detection: A discriminative regional feature integration approach. *International Journal of Computer Vision*, 123(2):251–268, 2017.

[144] Lijun Wang, Huchuan Lu, Xiang Ruan, and Ming-Hsuan Yang. Deep networks for saliency detection via local estimation and global search. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3183–3192, 2015.

[145] Lijun Wang, Huchuan Lu, Yifan Wang, Mengyang Feng, Dong Wang, Baocai Yin, and Xiang Ruan. Learning to detect salient objects with image-level supervision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3796–3805, 2017.

[146] Linzhao Wang, Lijun Wang, Huchuan Lu, Pingping Zhang, and Xiang Ruan. Saliency detection with recurrent fully convolutional networks. In *Proceedings of the European Conference on Computer Vision*, pages 825–841, 2016.

[147] Shenlong Wang, Sanja Fidler, and Raquel Urtasun. Proximal deep structured models. In *Proceedings of the Advances in Neural Information Processing Systems*, pages 865–873, 2016.

[148] Lina Wei, Shanshan Zhao, Omar El Farouk Bourahla, Xi Li, and Fei Wu. Group-wise deep co-saliency detection. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 3041–3047, 2017.

[149] Xiu-Shen Wei, Chen-Lin Zhang, Yao Li, Chen-Wei Xie, Jianxin Wu, Chunhua Shen, and Zhi-Hua Zhou. Deep descriptor transforming for image co-localization. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 3048–3054, 2017.

[150] Xiu-Shen Wei, Chen-Lin Zhang, Jianxin Wu, Chunhua Shen, and Zhi-Hua Zhou. Unsupervised object discovery and co-localization by deep descriptor transformation. *Pattern Recognition*, 88:113–126, 2019.

[151] Yunchao Wei, Jiashi Feng, Xiaodan Liang, Ming-Ming Cheng, Yao Zhao, and Shuicheng Yan. Object region mining with adversarial erasing: A simple classification to semantic segmentation approach. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6488–6496, 2017.

[152] Yunchao Wei, Xiaodan Liang, Yunpeng Chen, Xiaohui Shen, Ming-Ming Cheng, Jiashi Feng, Yao Zhao, and Shuicheng Yan. STC: A simple to complex framework for weakly-supervised semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(11):2314–2320, 2017.

[153] John M. Winn, Antonio Criminisi, and Thomas P. Minka. Object categorization by learned universal visual dictionary. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1800–1807, 2005.

[154] Chuan Yang, Lihe Zhang, Huchuan Lu, Xiang Ruan, and Ming-Hsuan Yang. Saliency detection via graph-based manifold ranking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3166–3173, 2013.

[155] Jimei Yang and Ming-Hsuan Yang. Top-down visual saliency via joint CRF and dictionary learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2296–2303, 2012.

[156] Xiwen Yao, Junwei Han, Dingwen Zhang, and Feiping Nie. Revisiting co-saliency detection: A novel approach based on two-stage multi-view spectral rotation co-clustering. *IEEE Transactions on Image Processing*, 26(7):3196–3209, 2017.

[157] Jason J. Yu, Adam W. Harley, and Konstantinos G. Derpanis. Back to basics: Unsupervised learning of optical flow via brightness constancy and motion smoothness. In *Workshops of the European Conference on Computer Vision*, pages 3–10, 2016.

[158] Ze-Huan Yuan, Tong Lu, and Yirui Wu. Deep-dense conditional random fields for object co-segmentation. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 3371–3377, 2017.

[159] Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *Proceedings of the European Conference on Computer Vision*, pages 818–833, 2014.

[160] Dingwen Zhang, Junwei Han, Jungong Han, and Ling Shao. Cosaliency detection based on intrasaliency prior transfer and deep intersaliency mining. *IEEE Transactions on Neural Networks and Learning Systems*, 27(6):1163–1176, 2016.

[161] Dingwen Zhang, Junwei Han, Chao Li, Jingdong Wang, and Xuelong Li. Detection of co-salient objects by looking deep and wide. *International Journal of Computer Vision*, 120(2):215–232, 2016.

[162] Dingwen Zhang, Junwei Han, and Yu Zhang. Supervision by fusion: Towards unsupervised learning of deep salient object detector. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4068–4076, 2017.

[163] Dingwen Zhang, Deyu Meng, and Junwei Han. Co-saliency detection via a self-paced multiple-instance learning framework. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(5):865–878, 2017.

[164] Jianming Zhang, Zhe L. Lin, Jonathan Brandt, Xiaohui Shen, and Stan Sclaroff. Top-down neural attention by excitation backprop. In *Proceedings of the European Conference on Computer Vision*, pages 543–559, 2016.

[165] Jianming Zhang, Stan Sclaroff, Zhe L. Lin, Xiaohui Shen, Brian L. Price, and Radomír Mech. Minimum barrier salient object detection at 80 FPS. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1404–1412, 2015.

[166] Pingping Zhang, Dong Wang, Huchuan Lu, Hongyu Wang, and Xiang Ruan. Amulet: Aggregating multi-level convolutional features for salient object detec-

tion. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 202–211, 2017.

[167] Pingping Zhang, Dong Wang, Huchuan Lu, Hongyu Wang, and Baocai Yin. Learning uncertain convolutional features for accurate saliency detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 212–221, 2017.

[168] Shuai Zheng, Sadeep Jayasumana, Bernardino Romera-Paredes, Vibhav Vineet, Zhizhong Su, Dalong Du, Chang Huang, and Philip H. S. Torr. Conditional random fields as recurrent neural networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1529–1537, 2015.

[169] Bolei Zhou, Aditya Khosla, Àgata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2921–2929, 2016.

[170] Chao Zhou, Hong Zhang, Xiaoyong Shen, and Jiaya Jia. Unsupervised learning of stereo matching. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1576–1584, 2017.

[171] Yanzhao Zhou, Yi Zhu, Qixiang Ye, Qiang Qiu, and Jianbin Jiao. Weakly supervised instance segmentation using class peak response. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3791–3800, 2018.

[172] Lingling Zhu, Zhibo Chen, Xiaoming Chen, and Ning Liao. Saliency & structure preserving multi-operator image retargeting. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1706–1710, 2016.