Faculty of Engineering and Information Technology
University of Technology Sydney

Image Co-saliency Detection and Co-segmentation from The Perspective of Commonalities

A thesis submitted in partial fulfillment of
the requirements for the degree of
Doctor of Philosophy

by

Lina Li

March 2020

CERTIFICATE OF AUTHORSHIP/ORIGINALITY

I certify that the work in this thesis has not previously been submitted for a degree nor has it been submitted as part of requirements for a degree except as fully acknowledged within the text.

I also certify that the thesis has been written by me. Any help that I have received in my research work and the preparation of the thesis itself has been acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

This thesis is the result of me conducted jointly with Shanghai University as part of a collaborative Doctoral degree.

This research is supported by the Australian Government Research Training Program.

Production Note:
Signature removed prior to publication.

Signature of Candidate: _____

Date: 2020.03.09

Acknowledgments

Foremost, I would like to express the deepest appreciation to my supervisor, Associate Prof. Jian Zhang, for the constant support and guidance for my Ph.D work. As a collaborative degree candidate, I would also like to thank Prof. Zhi Liu, my co-supervisor, for his patient guidance, and scientific advice. Without their generous support, this dissertation would not have been possible.

I would like to thank my labmates in Global Big Data Technology Center: Junjie Zhang, Yifan Zuo, Xiaoshui Huang, Muming Zhao, Hao Cheng, Lu Zhang, Yongshun Gong, and Zhibin Li and labmates in Shanghai University, for the help and discussions and for all the fun we have had in my PhD study period.

Last but not the least, many thanks would go to my parents and my sisters, for their support and confidence in me through all these years.

Contents

Certificate . . . . . **i**
Acknowledgment . . . . . **ii**
List of Figures . . . . . **vi**
List of Tables . . . . . **xii**
List of Publications . . . . . **xiii**
Abstract . . . . . **xv**

Chapter 1 Introduction **1**
1.1 Background 1
1.2 Research Problems 7
1.2.1 Global consistency among all images 7
1.2.2 Co-saliency fusion and refinement 7
1.2.3 Simple image selection 8
1.2.4 Simple image guidance 8
1.2.5 Common features in CNNs 9
1.2.6 The simplification of the CNNs model 9
1.3 Research Contributions 10
1.4 Thesis Structure 11

Chapter 2 Relevant Theories and Related Work **12**
2.1 Image Co-saliency Detection 12

2.1.1 Concepts 12
2.1.2 Image saliency detection 14
2.1.3 Image co-saliency detection 20
2.2 Image Co-segmentation 25
2.2.1 Concepts 25
2.2.2 Image segmentation 26
2.2.3 Image co-segmentation 31
2.3 Summary 35

Chapter 3 Co-saliency Detection Based on Region-level Fusion and Pixel-level Refinement 36
3.1 Motivations 36
3.2 Co-saliency Detection Model 39
3.2.1 Intra saliency 39
3.2.2 Region-level co-saliency 41
3.2.3 Pixel-level co-saliency 43
3.3 Experiments 45
3.4 Summary 54

Chapter 4 Unsupervised Image Co-segmentation via Guidance of Simple Images 55
4.1 Motivations and Significance 55
4.2 The Proposed Method 58
4.2.1 Image ranking 62
4.2.2 Simple image selection 67
4.2.3 Samples extraction 70
4.2.4 Complicated image segmentation 72
4.3 Experiments 74

4.3.1 Parameter setting 75
4.3.2 Results of the proposed method 75
4.3.3 Comparison with other methods 82
4.3.4 Discussion 85
4.4 Summary 86

Chapter 5 Learn Image Object Co-segmentation with Multi-scale Feature Fusion 87
5.1 Motivations and Significance 87
5.2 The Proposed Model 91
5.2.1 The reformed VGG network 92
5.2.2 Features extraction and fusion 93
5.2.3 Upsampling 94
5.3 Experiments 96
5.3.1 Experimental settings 96
5.3.2 Experimental results 97
5.4 Summary 99

Chapter 6 Conclusions and Future Work 100
6.1 Conclusions 100
6.2 Future Work 102

Bibliography 103

List of Figures

1.1 An example of a saliency map. From the left to right: (a) Original images, (b) the saliency map generated by an image object saliency detection model, (c) ground truth 1

1.2 From the left to right: (a) the simple images and its segmentation results, (b) complicated images. 2

1.3 Illustrations of the definition of objects and common objects. . 3

1.4 The saliency maps of complicated images. From top to bottom: (a) Original images, (b) the saliency map generated by an image object single saliency detection model, (c) ground truth. 5

2.1 An example of the saliency map, and the shifted eye fixation results in one early work. 13

2.2 One general architecture of early works from (Itti, Koch & Niebur 1998). 15

2.3 An image is over segmented into regions in Saliency Tree (Liu, Zou & Le Meur 2014). 16

2.4 An overview of FCN. 19

2.5 The flowchart of a typical bottom-up image co-saliency detection model (Fu, Cao & Tu 2013). 21

2.6 The flowchart of an image co-saliency detection model based on fusion of saliency maps (Cao, Tao, Zhang, Fu & Feng 2014). 23
2.7 The flowchart of an image co-saliency detection model based on CNNs (Wei, Zhao, Bourahla, Li, Wu & Zhuang 2019). . . . 25
2.8 A simple 2D segmentation example for a 3*3 image. The seeds are $O = v$ and $B = p$. The cost of each edge is reflected by the edges thickness. 27
2.9 Three examples of Grab-Cut. The user drags a rectangle loosely around an object. The object is then extracted automatically. 29
2.10 ASPP structure. 30
2.11 (a) Sparse feature extraction with standard convolution on a low resolution input feature map. (b) Dense feature extraction with atrous convolution with rate r = 2, applied on a high resolution input feature map. 31
2.12 The illustration of a deep CNNs for image object co-segmentation (Li, Jafari & Rother 2018). 35

3.1 Illustration of the proposed co-saliency model. From top to bottom: the image set (5 out of 36 images are shown), ultrametric contour maps, region segmentation results, intra saliency maps, initial region-level co-saliency maps, final region-level co-saliency maps, and pixel-level co-saliency maps. . . . 40

3.2 Co-saliency maps generated for image pairs in the CP dataset. From top to bottom: original image pairs, binary ground truths, co-saliency maps generated using Li's model (Li & Ngan 2011), Fu's model (Batra, Kowdle, Parikh, Luo & Chen 2010), Liu's model (Liu, Zou, Li, Shen & Le Meur 2013) and our model, respectively. 46

3.3 Co-saliency maps generated for image pairs in the CP dataset. From top to bottom: original image pairs, binary ground truths, co-saliency maps generated using Li's model (Li & Ngan 2011), Fu's model (Batra et al. 2010), Liu's model (Liu et al. 2013) and our model, respectively. 47

3.4 Co-saliency maps generated for an image set in the iCoseg dataset. From top to bottom in each sub-figure: some original images in the image set, binary ground truths, co-saliency maps generated using Fu's model (Batra et al. 2010), Liu's model (Liu et al. 2013) and our model, respectively. 48

3.5 Co-saliency maps generated for an image set in the iCoseg dataset. From top to bottom in each sub-figure: some original images in the image set, binary ground truths, co-saliency maps generated using Fu's model (Batra et al. 2010), Liu's model (Liu et al. 2013) and our model, respectively. 49

3.6 Co-saliency maps generated for an image set in the iCoseg dataset. From top to bottom in each sub-figure: some original images in the image set, binary ground truths, co-saliency maps generated using Fu's model (Batra et al. 2010), Liu's model (Liu et al. 2013) and our model, respectively. 50

3.7 Co-saliency maps generated for an image set in the iCoseg dataset. From top to bottom in each sub-figure: some original images in the image set, binary ground truths, co-saliency maps generated using Fu's model (Batra et al. 2010), Liu's model (Liu et al. 2013) and our model, respectively. 51
3.8 Precision-recall curves for different saliency maps on the CP dataset. 52
3.9 Precision-recall curves for different saliency maps on the iCoseg dataset. 53

4.1 The flowchart. There are four steps in the proposed method: image ranking, simple image selection, samples extraction and complicated image segmentation. 59
4.2 The pre-processing. One image is over-segmented into hierarchical regions in three layers and is represented by two histograms: color histogram and dense SIFT histogram. 60
4.3 Region matching. For one region in layer 2 of image 1, its most similar region is searched through all the three layers in image 2. 62
4.4 The region matching results. The 1st row shows original images, the 1st column shows the regions hierarchically extracted from the three layers, and the 2nd-5th column show the corresponding best matched regions in other images. 63
4.5 Image Ranking. Images are ranked based on their saliency maps. 64

4.6 Image ranking results with different saliency models. The 1st, 2nd and 3rd row show the ranking results based on saliency maps generated using ST (Liu et al. 2014), HS (Yan, Xu, Shi & Jia 2013) and RBD (Zhu, Liang, Wei & Sun 2014), respectively. 67
4.7 Simple image selection. There are two boxes. Top box: the original ranked images and their corresponding single saliency maps. Bottom box: the selected simple images and their corresponding co-segmentation results. 68
4.8 Experimental results. In each image group, from top to bottom, RI: Ranked images, ISR: Initial segmentation results, and FSR: Final segmentation results. 76
4.9 Performance on iCoseg Dataset with Different Features. 77
4.10 Performance on iCoseg Dataset with Different Saliency Models 78
4.11 Experimental results. (a) the original images, from (b) to (d), the results of the proposed method, (Faktor & Irani 2013) and (Lee, Jang, Sim & Kim 2015), respectively. 80
4.12 Experimental results. (a) the original images, from (b) to (d), the results of the proposed method, (Faktor & Irani 2013) and (Lee et al. 2015), respectively. 81
4.13 Experimental results. (a) the original images, from (b) to (d), the results of the proposed method, (Faktor & Irani 2013) and (Lee et al. 2015), respectively. 82
4.14 Weighted F-measure on iCoseg Dataset 83
4.15 Intersection-over-Union(IoU) on iCoseg Dataset 83
4.16 F-measure on iCoseg Dataset 84

5.1 An overview of the proposed model. It can be viewed as three parts. The input images are first passed through the reformed VGG networks, which share exactly the same layer structure and parameters. Then, multi-scale [2,4,8,16] features are extracted at each layer (represented in colors and they are stacked according to the scales). They are fused both within each image (in grey) into intra-image features and across images (in colors) into inter-image features, and then intra-image and inter-image features are further fused at each scale, the features at different scales are represented in the corresponding colors. At last, the fused multi-scale features are summed up and upsampled to obtained the coarse co-segmentation results, and final co-segmentation results are obtained via Grabcut refinement. 88

5.2 Experimental results, objective evaluations. 97

5.3 Experimental results. (a) the original images, from (b) to (d), the results of (Li, Liu & Zhang 2018), the proposed model, and ground truth, respectively. 98

5.4 Experimental results. (a) the original images, from (b) to (d), the results of (Li, Liu & Zhang 2018), the proposed model, and ground truth, respectively. 98

List of Tables

4.1 Overall Performance on iCoseg Dataset 85

5.1 The implementation details of layers in 5.2.1. 92
5.2 The implementation details of the layers in Section 5.2.2. . . . 94
5.3 The implementation details of the layers in 5.2.3 95

List of Publications

Papers Published

- **Lina Li**, Zhi Liu, Wenbin Zou, Xiang Zhang, Olivier Le Meur(2014), Co-saliency detection based on region-level fusion and pixel-level refinement. *in* 'Proceedings of the International conference on Multimedia and Expo (ICME'14)', IEEE, pp. 1-6.

- **Lina Li**, Zhi Liu, Jian Zhang, (2018), Unsupervised image co-segmentation via guidance of simple images. *Neurocomputing*, vol. 275, pp. 1650-1661.

- **Lina Li**, Zhi Liu, Jian Zhang, Xiaofei Zhou, Learn Image Object Co-segmentation with Multi-scale Feature Fusion. *in* 'Visual Communications and Image Processing (VCIP'19)', IEEE, Accepted.

- Shuhua Luo, Zhi Liu, **Lina Li**, Xuemei Zou, Olivier Le Meur(2013), Efficient saliency detection using regional color and spatial information, *in* '4th European Workshop on Visual Information Processing (EUVIP'13)', pp. 184-189.

- Zhi Liu, Wenbin Zou, **Lina Li**, Liquan Shen, Olivier Le Meur(2014),Co-saliency detection based on hierarchical segmentation, *IEEE Signal Processing Letters*, vol. 21, no. 1, pp. 88-92.

- Linwei Ye, Zhi Liu, **Lina Li**(2015), Evaluation on fusion of saliency and objectness for salient object segmentation, *in* 'Proceedings of the 7th International Conference on Internet Multimedia Computing and Service(ICIMCS'15)', ACM, p.18.

- Linwei Ye, Zhi Liu, **Lina Li**, Liquan Shen, Cong Bai, Yang Wang(2017), Salient object segmentation via effective integration of saliency and objectness. *IEEE Transactions on Multimedia*, vol. 19, no. 8. pp. 1742-1756.

Abstract

Image co-saliency detection and image co-segmentation aim to identify the common salient objects and extract them in a group of images.

Image co-saliency detection and image co-segmentation are important for many content-based applications such as image retrieval, image editing, and content aware image/video compression. The image co-saliency detection and image co-segmentation are very close works. The most important part in these two works is the definition of the commonality of the common objects. Usually, common objects share similar low-level features, such as appearances, including colours, textures shapes, etc. as well as the high-level semantic features.

In this thesis, we explore the commonalities of the common objects in a group of images from low-level features and high-level features, and the way to achieve the commonalities and finally segment the common objects. Three main works are introduced, including an image co-saliency detection model and two image co-segmentation methods.

Firstly, an image co-saliency detection model based on region-level fusion and pixel-level refinement is proposed. The commonalities between the common objects are defined by the appearance similarities on the regions from all the images. It discovers the regions that are salient in each individual image as well as salient in the whole image group. Extensive experiments

on two benchmark datasets demonstrate that the proposed co-saliency model consistently outperforms the state-of-the-art co-saliency models in both subjective and objective evaluation.

Secondly, an unsupervised images co-segmentation method via guidance of simple images is proposed. The commonalities are still defined by handcrafted features on regions, colours and textures, but not calculated among regions from all the images. It takes advantages of the reliability of simple images, and successfully improves the performance. The experiments on the dataset demonstrate the outperformance and robustness of the proposed method.

Thirdly, a learned image co-segmentation model based on convolutional neural network with multi-scale feature fusion is proposed. The commonalities between objects are not defined by handcraft features but learned from the training data. When training a neural network with multiple input images simultaneously, the resource cost will increase rapidly with the inputs. To reduce the resource cost, reduced input size, less downsampling and dilation convolution are adopted in the proposed model. Experimental results on the public dataset demonstrate that the proposed model achieves a comparable performance to the state-of-the-art methods while the network has successfully gotten simplified and the resources cost is reduced.

Chapter 1

Introduction

1.1 Background

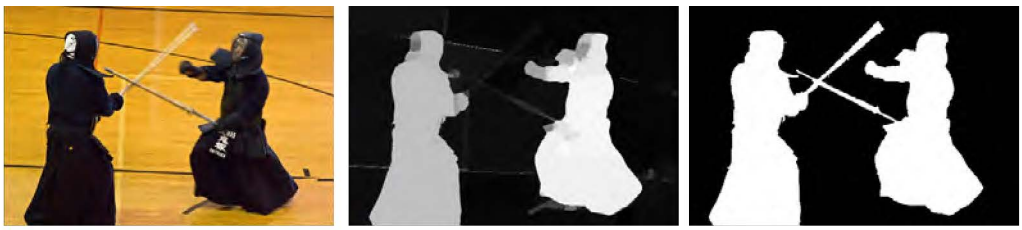

Figure 1.1: An example of a saliency map. From the left to right: (a) Original images, (b) the saliency map generated by an image object saliency detection model, (c) ground truth

The past twenty years have witnessed the booming research on saliency detection and a number of saliency models have been proposed and evaluated (Borji, Cheng, Jiang & Li 2015). The output of saliency model is saliency map, as shown in Figure 1.1, which indicates the importance of pixels/blocks/regions to human visual attention. Image object co-segmentation, like image co-saliency, has received a boosting attention by researchers who

work on object segmentation. The object saliency map as well as the object segmentation results serve as an indispensable cue for a variety of saliency-based applications such as human fixation prediction (Itti et al. 1998, Borji 2012, Borji & Itti 2012, Liu, Han, Zhang, Wen & Liu 2015), salient object detection (Borji et al. 2015), salient object segmentation (Liu, Shi, Shen, Xue, Ngan & Zhang 2012, Kruthiventi, Gudisa, Dholakiya & Venkatesh Babu 2016, Li, Xie, Lin & Yu 2017), content-aware image/video retargeting (Shamir & Sorkine 2009, Liang, Liu & Gutierrez 2016, Cho, Park, Oh, Tai & So Kweon 2017, Bansal, Ma, Ramanan & Sheikh 2018), and content-based image compression (Guo & Zhang 2009, Li, Zuo, Gu, Zhao & Zhang 2018), etc.

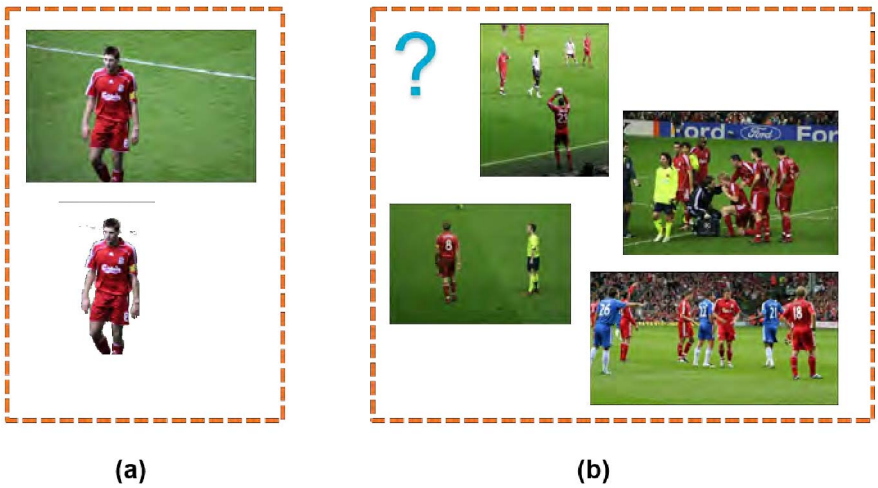

Figure 1.2: From the left to right: (a) the simple images and its segmentation results, (b) complicated images.

Generally, most saliency models are designed to detect salient objects in a single image, and it is difficult for the state-of-the-art single-image saliency models to effectively highlight salient objects in an image with a complex scene, as shown in Figure. 1.2. Similar to single object saliency detection, object segmentation in a single image usually incorporates different measures

at the level of pixels (Shi & Malik 2000, Boykov & Jolly 2001, Rother, Kolmogorov & Blake 2004) , or at the level of regions (Fulkerson, Vedaldi & Soatto 2009) into a global objective function, and by minimizing the objective function, the objects can be separated from the background, the object is segmented isolatedly within an image. This will be very difficult in some cases, for example, in Figure. 1.2, the images in box (b), variable objects are contained in these images and the backgrounds are cluttered, it is almost impossible for a single image object saliency detection model or a single image object segmentation method to accurately detect and segment the object that we need (the soccer players in red).

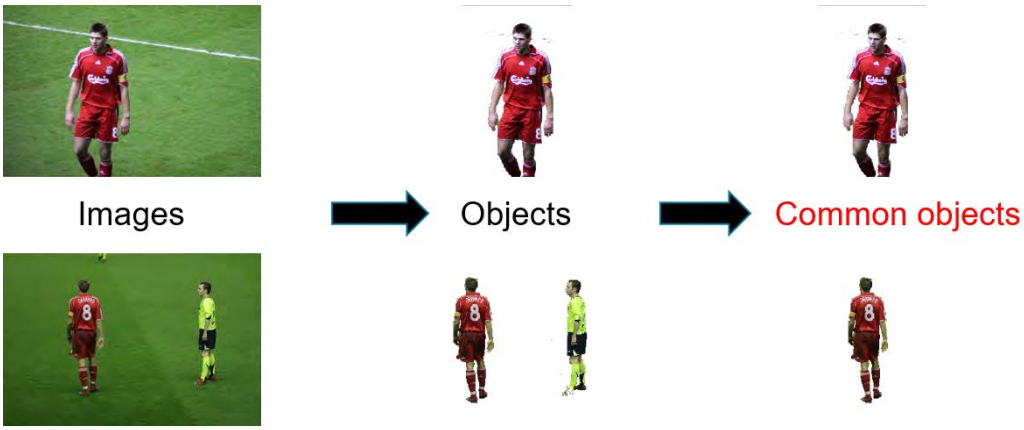

Figure 1.3: Illustrations of the definition of objects and common objects.

To solve the aforementioned problems, the detection and segmentation of common objects in a group of images, i.e. image co-saliency detection and image co-segmentation, attract more and more attention. Image co-saliency detection and co-segmentation models exploit the repeated occurrence of salient objects in a set of images to better highlight such co-salient objects and suppress irrelevant background regions than single-image saliency detection models. This kind of image co-saliency detection models and object co-segmentation methods aim to detect/segment objects with additional in-

formation from other images containing the same or similar objects (Rother, Minka, Blake & Kolmogorov 2006), as shown in Figure. 1.3, with the additional information from the images in box (a), the objects that we need are more obvious than that without them. Thus, image co-saliency detection and object co-segmentation can be derived as a task of jointly detecting/segmenting the common objects in a group of images.

To discover, detect and segment the common objects, first, we should define what is a common object. Usually, common objects share similar low-level features, such as appearances, including colors, textures shapes, etc. as well as the high-level semantic features. In some cases, common objects only share similar semantic features (e.g. dogs in different appearances and shapes are treated as the common objects).

At first, objects with similar appearances as the common objects, and the common features are handcrafted. Most of the unsupervised image co-saliency detection models and object co-segmentation methods utilize the correspondences among all the images, they treat the detected/segmented objects in all the images as the same, the objects been detected/segmented in most of the images are the common objects.

However, our further observation indicates that the object saliency detection or segmentation results in complicated images are often unsatisfactory, thus unreliable, as shown in Figure 1.4. Some of the co-saliency detection models and object co-segmentation methods select similar images first, then calculate the correspondences among the selected images, they may achieve satisfactory results for simple images; but not for complicated images as the detection/segmentation results in these complicated images are, again as we observed, not reliable. It could be quite difficult for these images with clutter background to distinguish objects from the background, and the correspon-

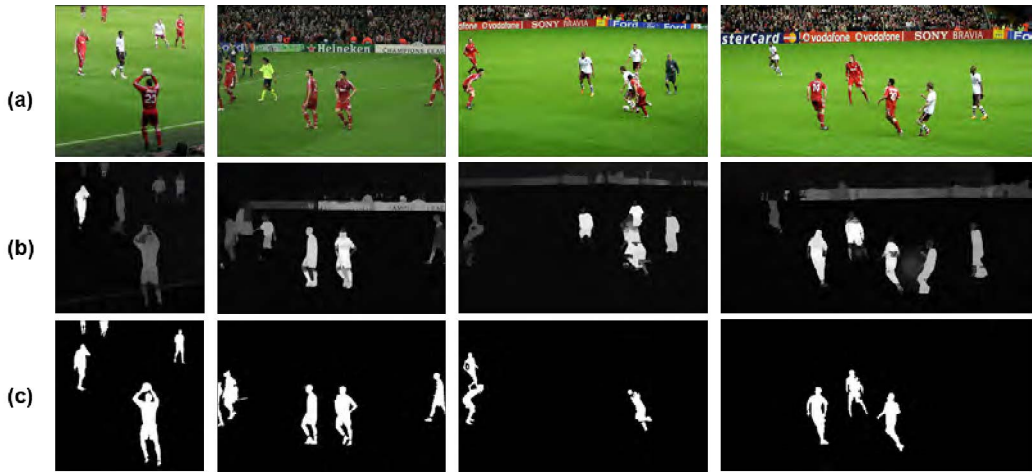

Figure 1.4: The saliency maps of complicated images. From top to bottom: (a) Original images, (b) the saliency map generated by an image object single saliency detection model, (c) ground truth.

dence calculation may be useless, even noisy. These limitations have been noticed in (Chen, Velasco-Forero, Tsang & Cham 2015), and the propagation method from simple images to complicated images has been used, however, this kind of propagation will lead that the failure results may also be propagated. Thus, it will be a good idea to explore the utilization of simple images, which may provides better information of common objects for complicated images. The simple images are to have better segmentation results which deserve more confidence; while the complicated images, on the contrary, need the aid of simple images, the complicated images are segmented with the guidance of simple images.

The past decade has witnessed the great development of deep neural network, and the power of deep features, especially in computer vision, such as image classification (Wang, Yang, Mao, Huang, Huang & Xu 2016, Wang, Jiang, Qian, Yang, Li, Zhang, Wang & Tang 2017), image saliency detection (Zhao, Ouyang, Li & Wang 2015, Zhang, Zhang, Dai, Harandi & Hartley

2018), image object detection/localization (Ren, He, Girshick & Sun 2015, Dai, Li, He & Sun 2016, Lin, Dollár, Girshick, He, Hariharan & Belongie 2017, Redmon, Divvala, Girshick & Farhadi 2016, Tompson, Goroshin, Jain, LeCun & Bregler 2015), image object segmentation (Oh, Benenson, Khoreva, Akata, Fritz & Schiele 2017, Badrinarayanan, Kendall & Cipolla 2017), and image semantic segmentation (Long, Shelhamer & Darrell 2015, Chen, Yang, Wang, Xu & Yuille 2016, Chen, Papandreou, Kokkinos, Murphy & Yuille 2018, Chen, Zhu, Papandreou, Schroff & Adam 2018), etc.

Some of the methods are unsupervised but utilizing not only the low-level features but also the high-level semantic features from some pre-trained Convolutional Neural Networks (CNNs) (Zhang, Han, Li & Wang 2015, Tsai, Li, Hsu, Qian & Lin 2018, Wei, Zhang, Wu, Shen & Zhou 2019).

Some methods are fully supervised and trained based on deep CNNs. In (Wei, Zhao, Bourahla, Li, Wu & Zhuang 2019), the networks are trained with shared branches to extract intra-image features for each input image as well as a combined branch to extract inter-image features (co-features), and the two features are combined to final segmentation predictions. In (Wei, Zhao, Bourahla, Li, Wu & Zhuang 2019), the input images are passed through a Siamese encoder to extract feature maps, then feature matching is performed using a mutual correlation network to obtain correspondence maps, and then, the common objects masks are obtained by passing concatenation of squeezed feature maps and correspondence maps through a Siamese decoder. When training a neural network with multiple input images simultaneously, the resource cost will increase rapidly with the inputs. Thus, reducing the cost while training a network for image co-segmentation is necessary.

1.2 Research Problems

Based on the aforementioned current research limitations, this thesis presents the following research problems.

1.2.1 Global consistency among all images

Some image co-saliency models are proposed to detect co-salient objects in a pair of images in the early work. Measuring the global consistency of an object in an image group is different. The salient objects should appear in most of the images frequently, the measurement of the frequency should be more general and effective for the whole image set. Also, the image co-saliency models that are calculated among multiple images, the computation cost will be much higher than the cost between image pairs, thus, the computation cost should be also considered.

Thus, to solve the image co-saliency detection problem, first, an effective scheme to measure the global consistency should be provided. Second, computation cost should be considered, e.g. computation based on regions/clusters (Fu et al. 2013) will be less than that based on pixels, for that regions can provide more local information than pixels and the amount of regions is much smaller than that of pixels.

Our solution is proposed and explained in Chapter 3, and details are discussed.

1.2.2 Co-saliency fusion and refinement

One key problem for an image co-saliency detection model in an image group to solve is how to fuse the global consistency (inter-image saliency) and the single image object saliency (intra-image saliency) into co-saliency. Some of

the image co-saliency models sum up or weighted sum up the two kinds of saliency maps together (Tan, Wan, Feng & Pun 2013, Rubinstein, Joulin, Kopf & Liu 2013), some of the models multiply them (Liu et al. 2013). And both of these two fusion schemes do not fuse the global consistency and intra-image saliency over the whole image set, i.e. the fusion of the two cues for an image is processed within the single image. To overcome the disadvantages in the fusion process, we fuse the two cues across the images, and refine the fusion results within each single image. Chapter 3 explains the details and analyses the experimental results.

1.2.3 Simple image selection

To select simple images that can guide the segmentation of the other images, two aspects should be considered. First, the simple images should be simple (only one or two objects in a simple image (Meng, Li, Ngan, Zeng & Wu 2013)), second, the objects detected in these simple images should be correct and accurate enough to guide the other images. In some image co-segmentation methods, simple images are ranked and selected only by their complexities (Chen et al. 2015) and (Meng et al. 2013), which ignore the global consistency of the objects in the simple images, and may select some noisy images as the guidance. To better select the simple images that easily segment the common objects, an effective image simplicity ranking method is proposed in Chapter 4. The detailed selection steps are explained and some results of the selection are also shown.

1.2.4 Simple image guidance

In (Chen et al. 2015), propagation method from simple images to complicated images has been used, but failure results may also be propagated.

And in (Meng et al. 2013), simple images are utilized to select the features and replace the fixed region similarity measurement for the better region matching. These two methods are not able to exclude the influence of the noisy complicated images. To overcome the disadvantages, an constant simple image guidance is presented in Chapter 4, as well as the details and the experimental results.

1.2.5 Common features in CNNs

Similar to some unsupervised methods, to be in accordance with the definition of the common objects, common features across images (inter-image saliency/segmentation) and single features within images (intra-image saliency/segmentation) should be fused together to co-segment the common objects. In (Wei, Zhao, Bourahla, Li, Wu & Zhuang 2019), the networks are trained with two shared branches to extract intra-image features for each input image as well as a combined branch to extract inter-image features, however, the features are not rich enough to gain. Chapter 5 proposes a multi-layer fusion CNN model to co-segment the objects, which extracts multi-layer multi-scale features within and across images.

1.2.6 The simplification of the CNNs model

When training a neural network with multiple input images simultaneously (Wei, Zhao, Bourahla, Li, Wu & Zhuang 2019), the cost of resources will increase rapidly with the inputs, which raises the requirements of hardware and lowers the training speed. When reducing the resource cost of a CNN, including the volume of the network and computation cost, it is usually considered to: 1) reduce the size of input images directly, and if there are down-sampling layers existed, that may lead some feature layers to the degeneration, espe-

cially the deeper layers (Krizhevsky, Sutskever & Hinton 2012, Simonyan & Zisserman 2015, He, Zhang, Ren & Sun 2016); 2) reduce the parameters, e.g. kernel sizes, then the performance of the network will be affected. To address the limitations we mention above, atrous convolution (Chen, Papandreou, Kokkinos, Murphy & Yuille 2018, Chen, Papandreou, Schroff & Adam 2017) is adopted in Chapter 5, which keeps the receptive field with less parameters, together with reduced input size and less downsampling layers.

1.3 Research Contributions

- An image co-saliency model by combining region-level fusion and pixel-level refinement is proposed. Using the two-level co-saliency measurement, the proposed model achieves a better co-saliency detection performance on two benchmark datasets compared to the state-of-the-art co-saliency models. (Chapter 3);

- A novel measurement and fusion scheme is proposed, which effectively exploits similarities between regions and global consistency over the image set, to fuse intra-image and inter-image cues over the image set into co-saliency; and exploits the similarities between pixel and region to refine the co-saliency cues into final co-saliency results. (Chapter 3);

- An effective framework to utilize simple images to segment complicated images is proposed, and the experiments on the iCoseg dataset demonstrate the outperformance and robustness of the proposed method. (Chapter 4);

- An effective method to rank image complexities via their saliency maps is proposed. (Chapter 4);

- A novel deep neural network with multi-scale feature fusion is proposed for image co-segmentation, and the experimental results demonstrate its comparable performance with the state-of-the-art methods. (Chapter 5);
- An effective scheme to simplify the image object co-segmentation network is proposed and the cost of resources is reduced while achieving comparable results. (Chapter 5).

1.4 Thesis Structure

The thesis is organized as follow:

Chapter 2 reviews the basic knowledge, relevant theories, and related work including image saliency detection, image object segmentation, image co-saliency detection/co-segmentation.

Chapter 3 first, briefly explains the motivations and significance of the proposed model, then introduces the details of the proposed co-saliency detection model, and then analyzes the experimental results.

Chapter 4 first, briefly explains the motivations and significance of the proposed method, then introduces the details of the proposed unsupervised image co-segmentation method, and then analyzes the experimental results.

Chapter 5 first, explains the motivations of the proposed CNN-based image co-segmentation model, then introduces the details of the proposed model, and then analyzes the experimental results.

Chapter 6 concludes the thesis and provides a discussion for future work.

Chapter 2

Relevant Theories and Related Work

This chapter reviews two works, image co-saliency detection and image object co-segmentation, in Section 2.1 and Section 2.2, respectively. Section 2.1 first introduces some concepts, then presents the basic and close related work, image saliency detection, and then reviews the works of image co-saliency detection. Section 2.2 first introduces some concepts, then presents the close related work, image object segmentation, and then reviews the works of image object co-segmentation. Finally, Section 2.3 ends this chapter with a summary.

2.1 Image Co-saliency Detection

2.1.1 Concepts

Humans are able to effortlessly and rapidly analyze a scene, and search the objects/regions that we are interested, (i.e., bottom-up, task-independent

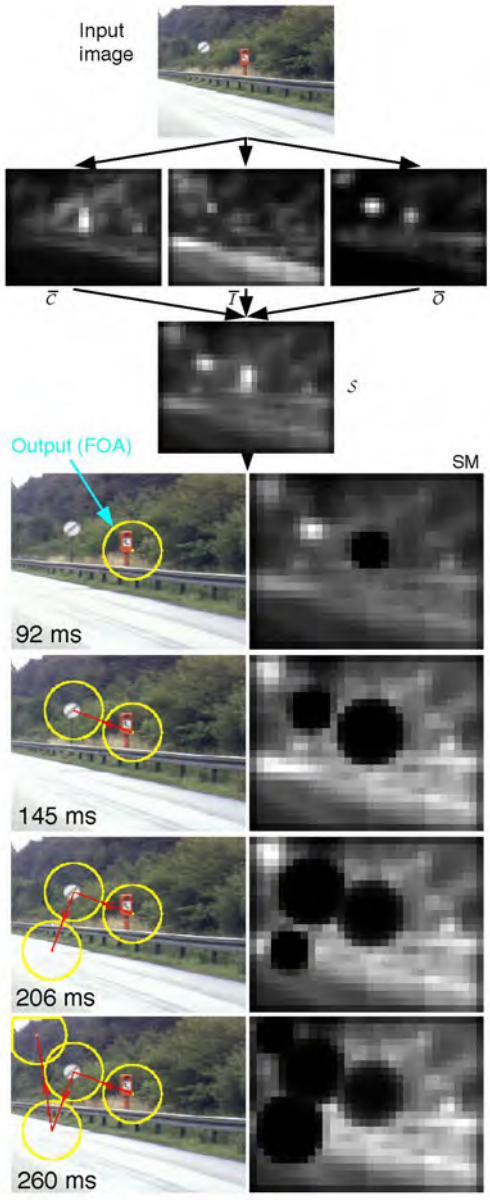

Figure 2.1: An example of the saliency map, and the shifted eye fixation results in one early work.

stage). These object/regions are then processed further in details to extract the high-level information (i.e., top-down, task-dependent stage). This capability has long been studied by cognitive scientists and has attracted a lot of interest in computer vision. Among the works of image saliency detection over the past decades, two concepts can be referred to saliency: fixation prediction (Itti et al. 1998, Bruce & Tsotsos 2006, Hou & Zhang 2007, Judd, Ehinger, Durand & Torralba 2009, Borji 2012, Borji & Itti 2012) and salient object detection (Goferman, Zelnik-Manor & Tal 2011, Margolin, Tal & Zelnik-Manor 2013, Li, Li, Shen, Dick & Van Den Hengel 2013, Yan et al. 2013, Zhang, Wang, Lu, Wang & Yin 2017, Wang, Borji, Zhang, Zhang & Lu 2017). Fixation prediction models try to predict where humans look, and output a small set of fixation points, while salient object detection models aim to detect the salient objects in a scene as completely as possible. In this thesis, image saliency detection that we mention (except in 2.1.2) refers to salient object detection. Image co-saliency detection, thus, refers to co-salient object detection. In this section, we first introduce the basic research of image co-saliency detection: image co-saliency detection.

2.1.2 Image saliency detection

Unsupervised saliency detection

One of the earliest work proposed a saliency model with multiple disciplines inspired by cognitive psychology, neuroscience, and computer vision (Itti et al. 1998). This model works at the bottom-up attention stage only and implements based on the center-surround mechanisms. The output of this model is the fixation prediction results, it uses 'Winner-take-all' to predict the locations that human eyes are interested, as well as the sequence that

the locations are shifted in, as shown in Figure 2.1.

There are also some works recently (Bruce & Tsotsos 2006, Hou & Zhang 2007, Judd et al. 2009, Borji 2012, Borji & Itti 2012) researching on fixation prediction. In (Borji 2012), low-level features such as orientation, color, intensity, saliency maps of previous best bottom-up models and top-down cognitive visual features (e.g., faces, humans, cars, etc.) are combined to train classifiers for predicting eye fixations. In (Borji & Itti 2012), local and global patch rarity color spaces are computed and fused to predict fixation locations.

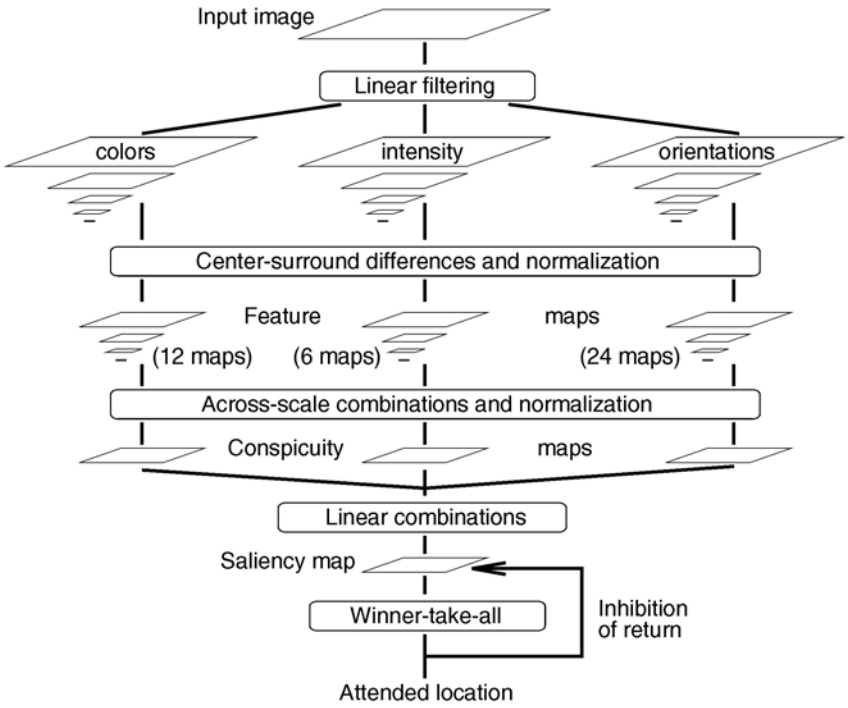

Figure 2.2: One general architecture of early works from (Itti et al. 1998).

In most of the works of saliency detection, saliency detection refers to salient object detection. They utilize various cues to distinguish the salient object from the background and noisy objects.

Early works extract features and computed on pixel-level. In (Itti et al. 1998), colours, intensity and orientations are used as the cues when compute pixel-wise center-around contrast, as shown in Figure 2.2. In (Hou & Zhang 2007, Zhou, Hou & Zhang 2010), (Zhang, Tong, Marks, Shan & Cottrell 2008) and (Harel, Koch & Perona 2007), the salient object is detected from the perspectives of spectral residual approach, information theory, and graph theory, respectively. In (Hu, Rajan & Chia 2005) and (Vidal, Ma & Sastry 2005), salient object is detected by estimating the subspaces of the features. In (Rosin 2009), an efficient and simple approach for salient object detection is proposed, which is parameter-free and needs only simple pixel-wise operations, e.g edge detection, threshold decomposition and moment preserving binarization. In (Achanta, Estrada, Wils & Süsstrunk 2008), a frequency-tuned approach is adopted, and saliency detection is defined as a binary segmentation issue. Multi-scale contrast is adopted in (Liu & Gleicher 2006, Liu, Yuan, Sun, Wang, Zheng, Tang & Shum 2010) to enhance the saliency maps.

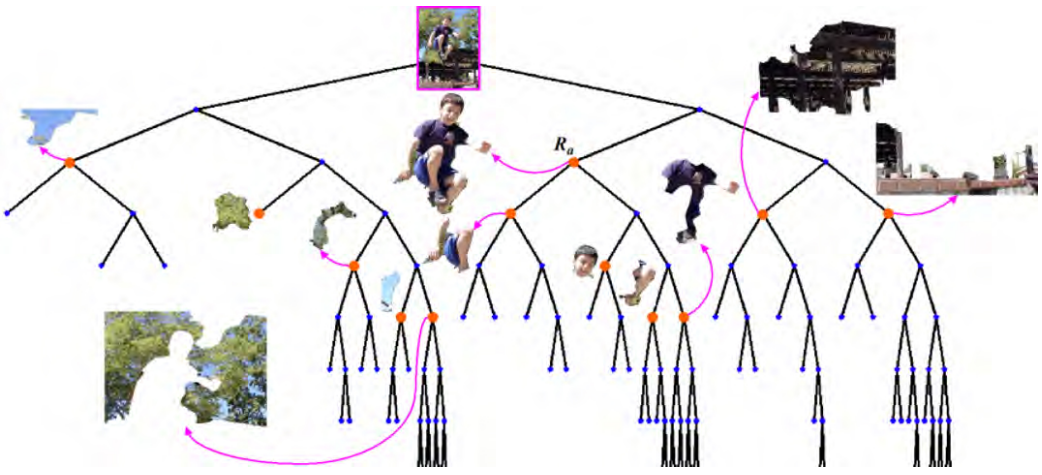

Figure 2.3: An image is over segmented into regions in Saliency Tree (Liu et al. 2014).

To better represent and obtain richer features of salient objects, some works represent objects and extract features and compute the contrast to the patch-level (Achanta et al. 2008, Klein & Frintrop 2011, Li, Li, Shen, Dick & Van Den Hengel 2013, Liu et al. 2010) and region-level (Liu & Gleicher 2006, Yu & Wong 2007, Cheng, Mitra, Huang, Torr & Hu 2014, Cheng, Warrell, Lin, Zheng, Vineet & Crook 2013), the regions may be obtained using methods, such as graph-based segmentation (Felzenszwalb & Huttenlocher 2004), mean-shift (Comaniciu & Meer 2002), SLIC (Achanta, Shaji, Smith, Lucchi, Fua & Süsstrunk 2012) or Turbopixels (Levinshtein, Stere, Kutulakos, Fleet, Dickinson & Siddiqi 2009), as shown in Figure 2.3 from (Liu et al. 2014), images are over-segmented into regions. For example, the contrast is measured by Kullback-Leibler divergence between distribution of features (Klein & Frintrop 2011). Despite local center-surround contrast, global contrast is also used to measure the uniqueness in many saliency models (Goferman et al. 2011, Margolin et al. 2013, Cheng, Mitra, Huang, Torr & Hu 2014, Cheng et al. 2013, Jiang & Davis 2013, Scharfenberger, Wong, Fergani, Zelek & Clausi 2013, Yan et al. 2013). In (Margolin et al. 2013), the uniqueness of a patch is measured by its distance to the average patch based on the observation that distinct patches are more scattered than non-distinct ones in the high-dimensional space. In (Cheng, Mitra, Huang, Torr & Hu 2014), the global contrast is computed between the target region with respect to all other image regions. In (Liu et al. 2014), a saliency-directed region merging approach with dynamic scale control scheme is proposed. Salient objects, in terms of uniqueness, can also be defined as the sparse noises in a certain feature space (Peng, Li, Ji, Hu, Xiong & Lang 2013, Shen & Wu 2012, Zou, Kpalma, Liu & Ronsin 2013). In (Zou et al. 2013), bottom-up segmentation is used as a guidance cue of the low-

rank matrix recovery for robustness purpose. In (Jiang, Wang, Yuan, Liu, Zheng & Li 2011), multi-scale local region contrast is proposed to enhance the robustness of the saliency map. Similar ideas of multi-hierarchical segmentations are also adopted in (Li, Li, Shen, Dick & Van Den Hengel 2013, Li, Lu, Zhang, Ruan & Yang 2013, Yan et al. 2013). In (Li, Li, Shen, Dick & Van Den Hengel 2013), a hypergraph, constructed by nonparametric multi-scale clustering of superpixels is built to obtain the internal consistency and external separation of regions.

Learning-based saliency detection

Recently, machine learning techniques, especially convolutional neural networks (CNNs) have been used to learn effective contextual representation for saliency detection.

Some models (Liu, Han, Zhang, Wen & Liu 2015, Li & Yu 2015, Zhao et al. 2015) adopt CNNs to extract features. In (Liu, Han, Zhang, Wen & Liu 2015, Li & Yu 2015), CNNs are used to extract multi-scale contextual features on multi-scale image regions for pixels and superpixels. In (Zhao et al. 2015), CNNs are used to learn both global and local contexts.

Some models (Kuen, Wang & Wang 2016, Li & Yu 2016, Liu & Han 2018, Liu & Han 2016, Wang, Wang, Lu, Zhang & Ruan 2016, Hou, Cheng, Hu, Borji, Tu & Torr 2017, Luo, Mishra, Achkar, Eichel, Li & Jodoin 2017, Zhang, Wang, Lu, Wang & Ruan 2017, Zhang, Wang, Lu, Wang & Yin 2017, Wang, Borji, Zhang, Zhang & Lu 2017) adopt fully convolutional networks (FCNs) (Long et al. 2015) for feature representation at each image location and generate saliency map in a convolutional way. FCNs can efficiently learn to make dense predictions for per-pixel tasks like semantic segmentation, saliency prediction, and segmentation etc. as shown in Figure 2.4. In (Li &

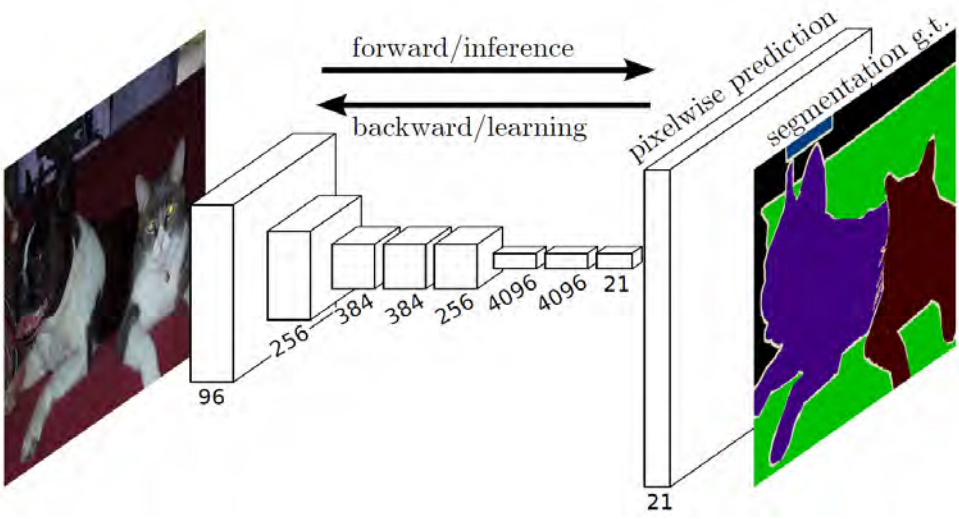

Figure 2.4: An overview of FCN.

Yu 2016), an FCN based saliency model and a multiscale image region based saliency model are combined. In (Wang, Wang, Lu, Zhang & Ruan 2016), FCN is recurrently used to refine saliency maps progressively. In (Liu & Han 2016), a U-Net based network is used to hierarchically predict and refine saliency maps from globally to locally. In (Luo et al. 2017, Zhang, Wang, Lu, Wang & Ruan 2017) U-Net based models are also used to incorporate multi-level contexts. In (Wang, Borji, Zhang, Zhang & Lu 2017), saliency maps are also refined by combining local and global context information. In (Hou et al. 2017), short connections are introduced into the multi-scale side outputs within the HED network (Xie & Tu 2015) to improve saliency detection performance. In (Hu, Shuai, Liu & Wang 2017), a loss based on level sets is adopted to train their saliency detection network and the saliency maps are refined using guided superpixel filtering. PiCANets (Liu, Han & Yang 2018) learns to selectively attend to informative context locations for each pixel, based on the observation that not all contextual information is helpful for the final task.

2.1.3 Image co-saliency detection

The image co-saliency detection can be viewed to be based on image saliency detection. Early works of image co-saliency detection models were bottom-up, unsupervised. With the rapid development of machine learning, especially the development of deep learning, CNNs-based top-down co-saliency detection models have been proposed and are rising up.

Unsupervised co-saliency detection

The problem of bottom-up based co-saliency detection usually can be formulated as

$$Cosaliency = Saliency \cdot Frequency. \tag{2.1}$$

Some of the models intergrated the intra-saliency cues with inter-saliency cues into co-saliency map. They use handcrafted features to score each pixels/regions in the image group (Fu et al. 2013, Ge, Fu, Liu, Bai & Yang 2016, Jing, He, Han & Niu 2015, Li & Ngan 2011, Lou, Xu, Xia, Yang & Ren 2017, Shen, Li, Zhou, Huo & Zhuang 2016, Zhu, Ji, Jiang, Xu & Liu 2017, Cong, Lei, Fu, Huang, Cao & Hou 2017), as shown in Figure 2.5 from (Fu et al. 2013), which mainly consist of feature extraction, single-image and multi-image cue exploration, and saliency map combination stages.

Some co-saliency models are proposed to detect co-salient objects in a pair of images. In an early work (Jacobs, Goldman & Shechtman 2010), the local structure changes induced by salient objects between the two images, which are captured for nearly the same scene, are used for co-saliency detection. In (Chen 2010), a preattentive scheme which utilizes the joint information between image pairs is exploited to detect co-salient objects. In (Rubinstein et al. 2013), the weighted Gist and SIFT flow are used to compute correspondences between images for facilitating the segmentation of common

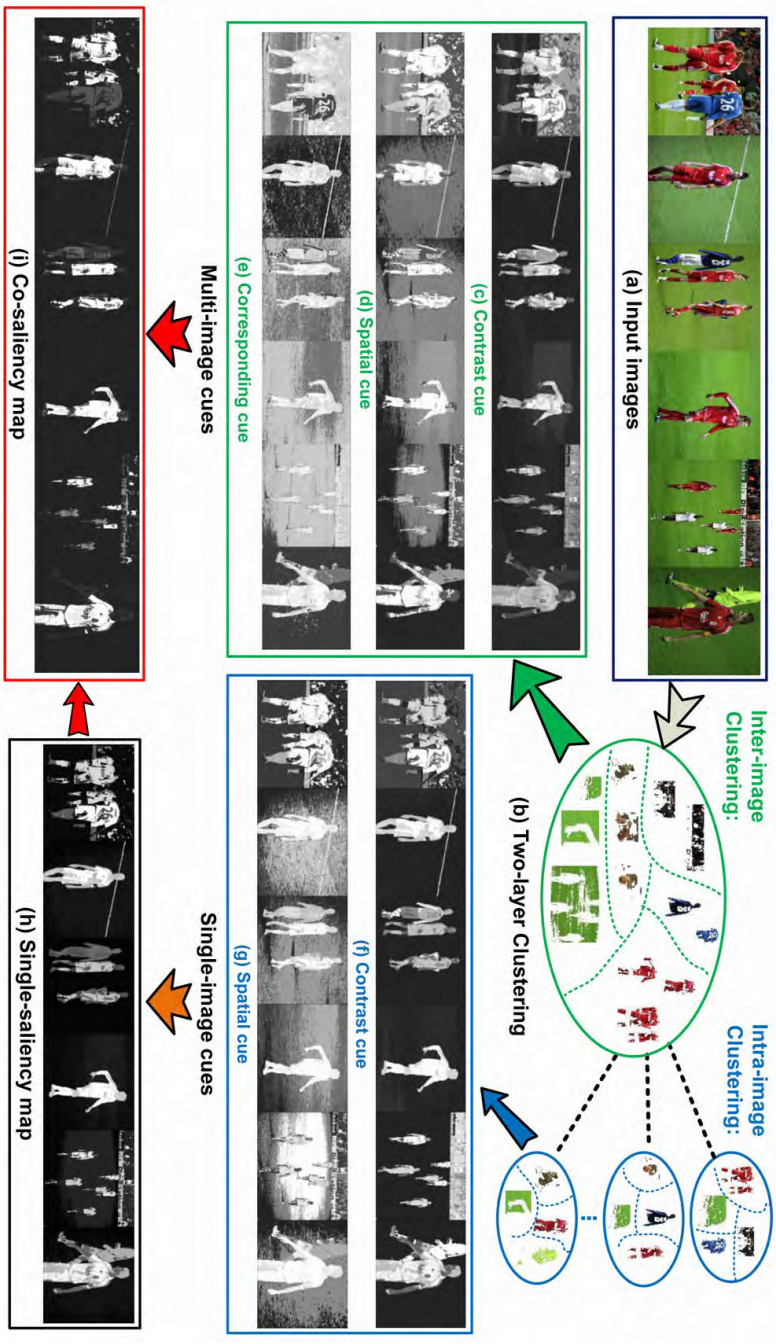

Figure 2.5: The flowchart of a typical bottom-up image co-saliency detection model (Fu et al. 2013).

objects in internet images. In (Zhu, Cai, Zheng, Wu & Thalmann 2013), with the aid of retrieving similar images using Google, object frequency and foreground likelihood are estimated for segmenting salient objects in the user input image. In (Li & Ngan 2011), co-saliency is modeled as a linear combination of the single image saliency map, which is computed using three available single-image saliency models, and multi-image saliency based on a co-multilayer graph. In (Tan et al. 2013), superpixel affinity matrix for image pair is calculated and propagated using SimRank algorithm to derive co-saliency maps. In (Fu et al. 2013), the cluster-based co-saliency model first computes the cluster-level co-saliency measures by integrating the contrast cue, spatial cue and corresponding cue among different images, and generates pixel-level co-saliency maps based on the likelihoods of pixels belonging to clusters. In (Liu et al. 2013), the hierarchical segmentation based co-saliency model measures regional contrast and similarity between regions in different images on the fine segmentation, evaluates object prior of regions on the coarse segmentation, and integrates them to generate the co-saliency map.

Some image co-saliency detection models fuse existing single-saliency/co-saliency maps into final co-saliency map, as shown in Figure 2.6. Given a group of images with similar objects, it first utilizes saliency detection algorithms to generate a group of saliency maps for all the images. The matrix jointing these feature histograms appears low rank, as the feature representation of the co-salient regions should be both similar and consistent. The general consistency criterion is formalized as the rank constraint, two consistency energies are proposed to describe it, which are based on low-rank matrix approximation and low-rank matrix recovery, respectively. By calculating the self-adaptive weight based on the consistency energy, the common

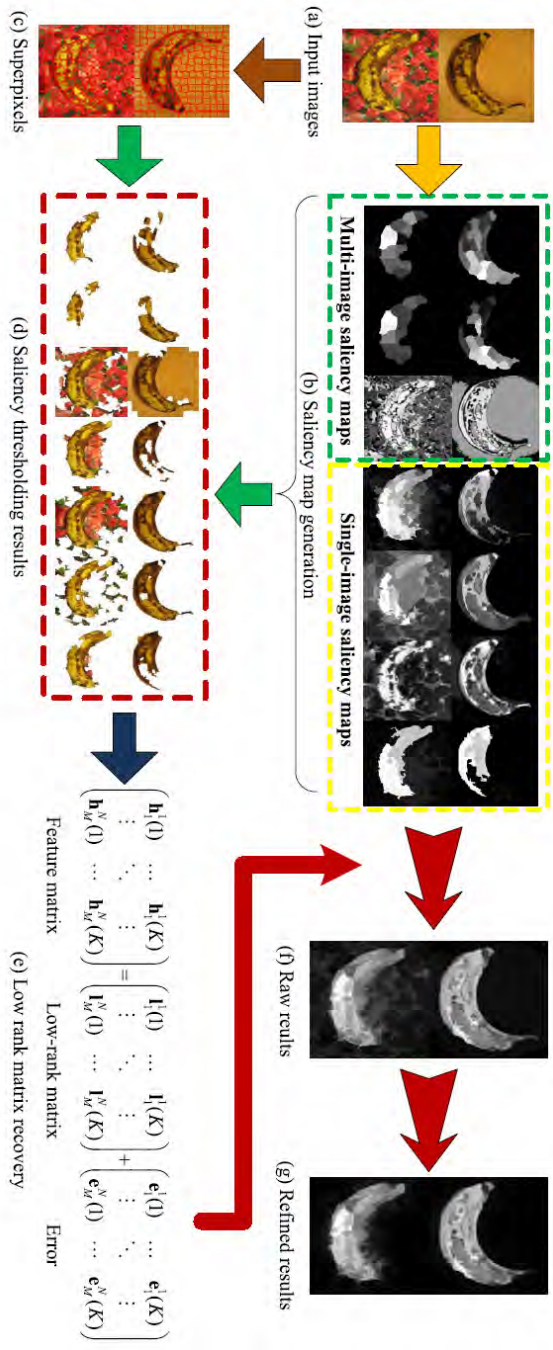

Figure 2.6: The flowchart of an image co-saliency detection model based on fusion of saliency maps (Cao, Tao, Zhang, Fu & Feng 2014).

salient regions are highlighted. In (Cao, Tao, Zhang, Fu & Li 2013), low-rank decomposition is applied to exploit the relationship of generated maps from multiple existing saliency and co-saliency approaches to obtain self-adaptive weights and then used these weights to combine the maps to generate the final co-saliency map. A reconstruction-based fusion approach is proposed in (Cao, Cheng, Tao & Fu 2014) which combines the knowledge mined from several existing saliency detection methods based on the reconstruction errors. Multiscale saliency maps are fused by using low-rank analysis and introducing a GMM-based co-saliency prior (Huang, Feng & Sun 2015).

Learning-based co-saliency detection

Learning-based co-saliency detection models aim at learning the patterns of the co-salient objects from a given image group (Zhang, Meng & Han 2016, Han, Cheng, Li & Zhang 2017). In (Cheng, Mitra, Huang & Hu 2014), correlations across Internet images belonging to the same category are exploited by building coarse appearance models for object regions and backgrounds and then improving saliency detection and image segmentation using these appearance models. In (Zhang et al. 2016, Zhang, Meng, Li, Jiang, Zhao & Han 2015) a self-paced multiple-instance learning (SP-MIL) framework is proposed to gradually learn the faithful patterns of the co-salient objects from confident image regions to ambiguous ones. In (Han et al. 2017),a unified metric learning-based framework to jointly learn discriminative feature representation and co-salient object detector is proposed in which the object function explicitly embeds a metric learning regularization term into support vector machine (SVM) training.

To utilize to powerful ability of CNNs in feature representation, some CNNs-based co-saliency detection models are proposed (Wei, Zhao, Bourahla,

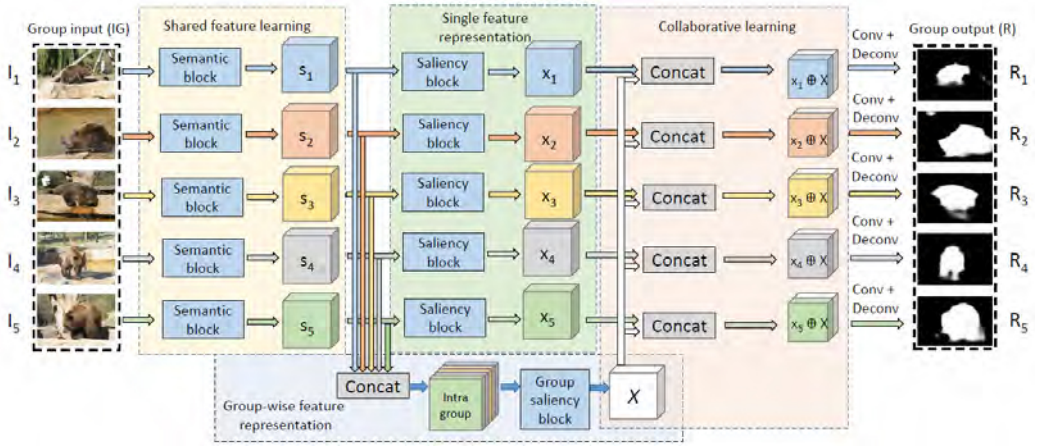

Figure 2.7: The flowchart of an image co-saliency detection model based on CNNs (Wei, Zhao, Bourahla, Li, Wu & Zhuang 2019).

Li, Wu & Zhuang 2019, Zhang, Han, Li & Wang 2015, Hsu, Tsai, Lin, Qian & Chuang 2018). High-level semantic features from CNNs are extracted in (Zhang, Han, Li & Wang 2015) to discover inter-image correspondence. An unsupervised CNNs (Hsu, Tsai, Lin, Qian & Chuang 2018) is proposed to jointly optimize the co-saliency maps. In (Wei, Zhao, Bourahla, Li, Wu & Zhuang 2019), an end-to-end framework based on the FCN (Long et al. 2015) is proposed to discover co-salient objects, as shown in Figure 2.7.

2.2 Image Co-segmentation

2.2.1 Concepts

Object co-segmentation is the task of segmenting the common objects from a set of images. Segmenting objects in an image is one of the fundamental tasks in computer vision. We first introduce the related work of image segmentation, then the related work of image co-segmentation.

2.2.2 Image segmentation

In general, image segmentation techniques have been divided into edge-based (Arbelaez, Maire, Fowlkes & Malik 2011), threshold-based (Cheriet, Said & Suen 1998, Hammouche, Diaf & Siarry 2008), region-based (Zhu & Yuille 1996, Freixenet, Muñoz, Raba, Martí & Cufí 2002), CNNs-based (Liu, Lin & Shen 2015, Badrinarayanan et al. 2017), clustering-based (Pappas 1992, Chuang, Tzeng, Chen, Wu & Chen 2006), wavelets-based (Porter & Canagarajah 1996, Choi & Baraniuk 2001), etc. Similar to image saliency detection, with the rapid development of machine leaning and deep learning, learning-based image segmentation methods have been proposed and received booming attentions. Due to the massive research works on image segmentation, we do not review all of them, we only introduce some most relevant related works, including graph-based segmentation methods and learning-based methods.

Graph-based image segmentation

Graph-based image segmentation techniques generally formulate the problem in terms of a graph $G = (V, E)$ where each node $v_i \in V$corresponds to a pixel in the image, and an edge $(v_i, v_j) \in E$connects vertices v_i and v_j , has a corresponding weight $\omega(v_i, v_j)$, which is a non-negative measure of the dissimilarity between neighboring elements v_i and v_j. There may or may not be an edge connecting each pair of vertices. In the case of image segmentation, the elements in V are pixels and the weight of an edge is some measure of the dissimilarity between the two pixels connected by that edge (e.g., the difference in intensity, color, motion, location or some other local attribute).

The earliest graph-based methods use fixed thresholds and local measures

in computing a segmentation. In (Zahn 1970), a segmentation method based on the minimum spanning tree (MST) of the graph is proposed in which the segmentation criterion is to break MST edges with large weights, however, differences between pixels within the high variability region can be larger than those between the ramp and the constant region. In (Wu & Leahy 1993), a segmentation method based on finding minimum cuts in a graph is proposed, where the cut criterion is designed in order to minimize the similarity between pixels that are being split, but it was biased toward finding small components. Then this bias was addressed with the *normalized cut* criterion developed by (Shi & Malik 2000). In (Felzenszwalb & Huttenlocher 2004), the problem of segmentation is addressed into regions, i.e. the elements in V are regions.

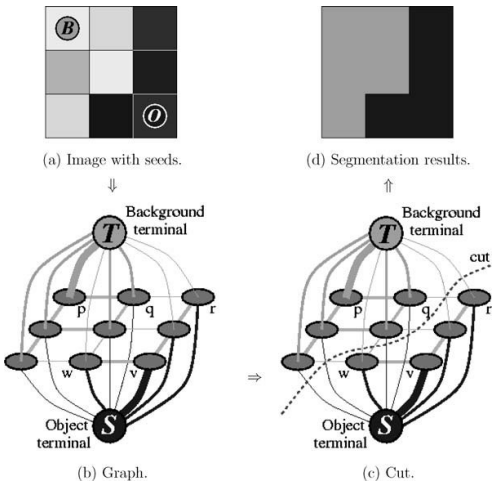

Figure 2.8: A simple 2D segmentation example for a 3*3 image. The seeds are $O = v$ and $B = p$. The cost of each edge is reflected by the edges thickness.

In (Boykov & Funka-Lea 2006), combinatorial graph-cut has been suc-

cessfully applied to segmentation of objects, as shown in Figure 2.8. It is based on s-t graph cut algorithm, and the problem was formulated as Maximum A Posterior estimation of a Markov Random Field (MAP-MRF) that required minimization of posterior energy

$$E(I) = -\lambda \Sigma_{p \in P} \ln Pr(I_p|I^o) + \Sigma_{p,q \in N} \delta_{I_p \neq I_q}, \tag{2.2}$$

where

$$\delta_{I_p \neq I_q} = \begin{cases} 1 & I_p \neq I_q \\ 0 & I_p = I_q. \end{cases}$$

is a Kronecker delta representing interaction potential, $I = I_p|p \in P$ is an unknown vector of original binary intensities $I_p \in 0, 1$ of image pixels P, vector I^o represents observed binary intensities corrupted by noise, and N is a set of all pairs of neighboring pixels. The nodes of the graphs represent image pixels/voxels. There are also two specially designated terminal nodes S (source) and T (sink) that represent object and background labels. Typically, neighboring pixels are interconnected by edges in a regular grid-like fashion. Edges between pixels are called n-links where n stands for neighbor. Note that a neighborhood system can be arbitrary and may include diagonal or any other kind of n-links. Another type of edges, called t-links, are used to connect pixels to terminals. All graph edges $e \in E$ including n-links and t-links are assigned some nonnegative weight (cost) w_e. An s-t cut is a subset of edges $e \in E$ such that the terminals S and T become completely separated on the induced graph $G(C) = \langle V, E|C \rangle$. Note that a cut(2.8(c))divides the nodes between the terminals. The cost of a cut is defined as the sum of the costs of edges that

$$|C| = \sum_{e \in E} w_e$$

The solution of optimizing the cost can be proved to be as the minimization of Equation 2.2.

In (Rother et al. 2004), an iterative graph-cut optimization procedure is adopted and used to simplify substantially the user interaction needed for a given quality of the result. Two enhancements are made to the graph cuts mechanism, iterative estimation, and incomplete labeling, which together allow a considerably reduced degree of user interaction for a given quality of result, as shown in Figure. 2.9, the only information needed to input is the box that contains the objects. (Tang, Gorelick, Veksler & Boykov 2013) proposed an energy term explicitly measuring the L1 distance between the object and background appearance models that can be globally maximized in one graph cut, which further improves the performance of Grab-Cut.



Figure 2.9: Three examples of Grab-Cut. The user drags a rectangle loosely around an object. The object is then extracted automatically.

Learning-based image segmentation

CNNs deployed in a fully convolutional manner (FCNs) have achieved remarkable performance on several semantic segmentation benchmarks.

Context information is significant for pixel labeling tasks (Li & Yu 2015, Liu et al. 2018, Noh, Hong & Han 2015, Dai, He & Sun 2016, Lin, Milan, Shen & Reid 2017, Chen et al. 2016, Chen, Zhu, Papandreou, Schroff & Adam 2018, Zhang, Dana, Shi, Zhang, Wang, Tyagi & Agrawal 2018). Therefore, PSPNet (Zhao, Shi, Qi, Wang & Jia 2017) performs spatial pyramid

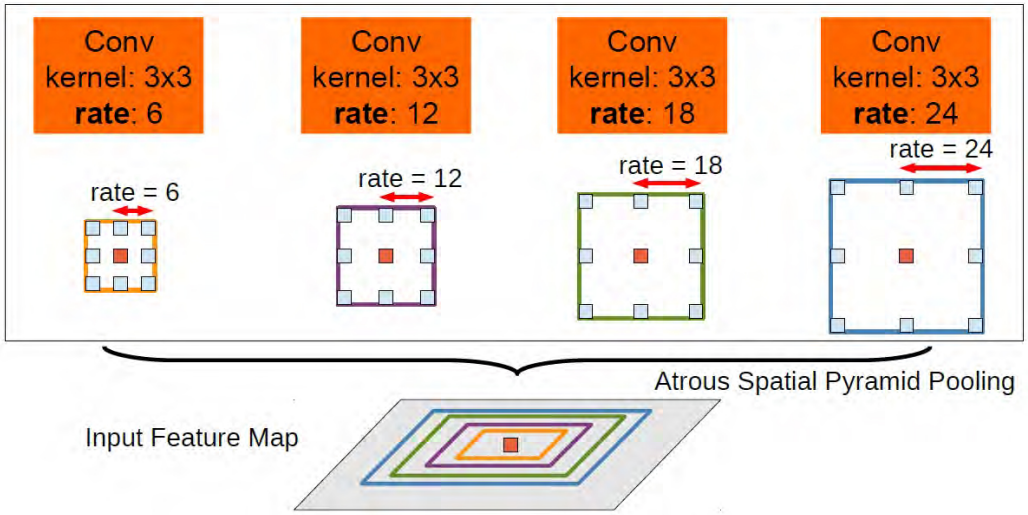

Figure 2.10: ASPP structure.

pooling (He, Zhang, Ren & Sun 2015) at several grid scales, while DeepLab (Chen, Papandreou, Kokkinos, Murphy & Yuille 2018, Chen et al. 2017) applies several parallel atrous convolution with different rates, as shown in Figure 2.10. Atrous convolution allows us to explicitly control the resolution at which feature responses are computed within Deep Convolutional Neural Networks, as shown in Figure 2.11. On the other hand, the improvement of neural network design has significantly improve the performance from AlexNet (Krizhevsky et al. 2012), VGG (Simonyan & Zisserman 2015), Inception (Szegedy, Vanhoucke, Ioffe, Shlens & Wojna 2016, Szegedy, Ioffe, Vanhoucke & Alemi 2017), ResNet (He et al. 2016) to more advanced architectures, such as Wide ResNet (Zagoruyko & Komodakis 2016), ResNeXt (Xie, Girshick, Dollár, Tu & He 2017), DenseNet (Huang, Liu, Van Der Maaten & Weinberger 2017), and Xception (Chollet 2017). The encoder-decoder structures (Ronneberger, Fischer & Brox 2015, Newell, Yang & Deng 2016, Badrinarayanan et al. 2017, Lin, Milan, Shen & Reid 2017, Pohlen, Hermans, Mathias & Leibe 2017, Amirul Islam, Rochan, Bruce & Wang 2017, Chen,

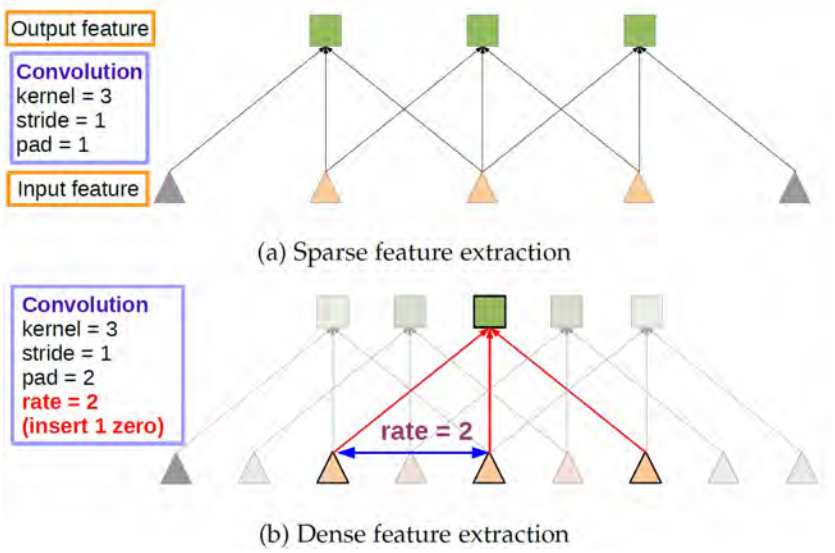

Figure 2.11: (a) Sparse feature extraction with standard convolution on a low resolution input feature map. (b) Dense feature extraction with atrous convolution with rate r = 2, applied on a high resolution input feature map.

Zhu, Papandreou, Schroff & Adam 2018, Zhang, Zhang, Peng, Xue & Sun 2018, Xiao, Liu, Zhou, Jiang & Sun 2018, Fu, Liu, Wang, Zhou, Wang & Lu 2019) can efficiently capture the long-range context information while keeping the detailed object boundaries. FRRN (Pohlen et al. 2017) employs a two-stream system, where full-resolution information is carried in one stream and context information in another. GridNet (Fourure, Emonet, Fromont, Muselet, Tremeau & Wolf 2017), with a similar idea, contains multiple streams with different resolutions.

2.2.3 Image co-segmentation

Unsupervised image co-segmentation

Most of the early co-segmentation methods are unsupervised and have various problem formulations (Li, Liu, Li, Lu & Ma 2016, Lin, Chen & Chang 2015,

Joulin, Bach & Ponce 2010, Rubinstein et al. 2013, Chen et al. 2015, Lee et al. 2015, Yu & Qi 2014, Faktor & Irani 2013, Meng et al. 2013, Zou, Liu, Kpalma, Ronsin, Zhao & Komodakis 2015). In (Joulin et al. 2010), the segmentation problem is handled in a discriminative clustering framework, in which objects are separated from the background by merging the image pixels into two clusters that can be maximally distinguished. Disturbance from the consistent background is considered in (Li et al. 2016), and a framework of saliency detection and discriminative learning is designed to overcome the disturbance. In (Lin et al. 2015), each image is treated as a player, for which a constrained utility function is designed by using the cooperative game, heat diffusion, and image saliency. Common objects can be segmented according to the labels defined via the constrained utility function. In (Rubinstein et al. 2013), pixel-level correspondences between images are considered, while in (Faktor & Irani 2013, Yu & Qi 2014), it is region-level affinities that are calculated. In (Faktor & Irani 2013), the affinities are calculated to score the multiple overlapping segment candidates, then these candidates and their scores are used to generate co-segmentation likelihood maps. In (Yu & Qi 2014), region-level affinities are utilized for each superpixel to find its correspondingly matched superpixels. Some co-segmentation methods utilize correspondences among all the images (Lee et al. 2015) while others compute correspondences among specific selected images (Rubinstein et al. 2013) . In (Chen et al. 2015), a feature pool is built based on overlapped rectangle regions and grows with the number of processed images, and the image to be processed obtains its prior information via the feature pool. In (Cho, Kwak, Schmid & Ponce 2015, Vo, Bach, Cho, Han, LeCun, Pérez & Ponce 2019), the problem is addressed as unsupervised discovery and localization of dominant objects from a noisy image collection with multiple object classes,

without even image-level annotations or any assumption of a single dominant class. In (Meng et al. 2013), a feature adaptive co-segmentation method is proposed to replace the fixed region similarity measurement for the better region matching. This method first evaluates the image complexities, then selects simple images and obtains the initial segments. This method uses a linear combination of the common features, and the linear combination parameters are learned from these segments. The learned adaptive feature is used to segment all the images in an image group. In (Zou et al. 2015), a saliency boosting model is proposed to optimize the joint segmentation results and saliency maps mutually.

Another kind of co-segmentation methods is interactive co-segmentation that allows a user to decide what the foreground is (Batra et al. 2010). This kind of methods may have simpler and reasonable energy functions by keeping a user in the loop, however, keeping a user involved is not always available for some applications.

Image co-segmentation has been extended from two-class segmentation to multi-class segmentation recently (Joulin, Bach & Ponce 2012, Kim & Xing 2012, Chang & Wang 2015, Yuan, Lu & Shivakumara 2014). In (Joulin et al. 2012), an energy based formulation of multi-class image co-segmentation is proposed which admits a probabilistic interpretation. An iterative scheme that alternates between a foreground modeling module and a region assignment module is proposed in (Kim & Xing 2012), which considers not only the multi-foreground segmentation but also the figure/ground assignment. Background regions can be automatically disregarded with the decomposition framework proposed in (Chang & Wang 2015) . In (Yuan et al. 2014), a scene image co-segmentation method is designed to segment images in the same scene into regions corresponding to their respective classes.

Learning-based image co-segmentation

In (Zhang, Han, Li & Wang 2015, Tsai, Li, Hsu, Qian & Lin 2019), the similarities between regions are computed both by traditional low-level features and high-level CNN features. In (Quan, Han, Zhang & Nie 2016), a manifold ranking algorithm for object co-segmentation is proposed by combining low-level appearance features and high-level semantic features which are pre-trained on the ImageNet dataset. In (Wei, Zhang, Wu, Shen & Zhou 2019), a Deep Descriptor Transformation (DDT) method is proposed to evaluate the correlations of deep descriptors generated from pre-trained models and then locate the common category-consistent object in a set of unlabeled images. In (Hsu, Lin & Chuang 2018), the common objects are optimized to get separated from the backgrounds, to achieve the highest similarities between the objects and lowest similarities between the objects and backgrounds, based on the high-level pre-trained CNNs features.

In (Vicente, Rother & Kolmogorov 2011), a pool of object-like proposals is first generated, then, a random forest classifier is trained to score the similarity between each pair of proposals. A deep dense conditional random field framework for object co-segmentation is proposed by inferring co-occurrence maps in (Yuan, Lu & Wu 2017). These co-occurrence maps measure the objectness scores, which are obtained using selective search (Uijlings, Van De Sande, Gevers & Smeulders 2013). Selective search uses hand-crafted SIFT and HOG features to generate object proposals. Thus, (Yuan et al. 2017) cannot be trained end-to-end.

In (Li, Jafari & Rother 2018), the method is trained end-to-end using a mutual correlation layer to detect the common objects. As shown in Figure 2.12, the input images are passed through a Siamese encoder to extract feature maps, then feature matching is performed using a mutual correla-

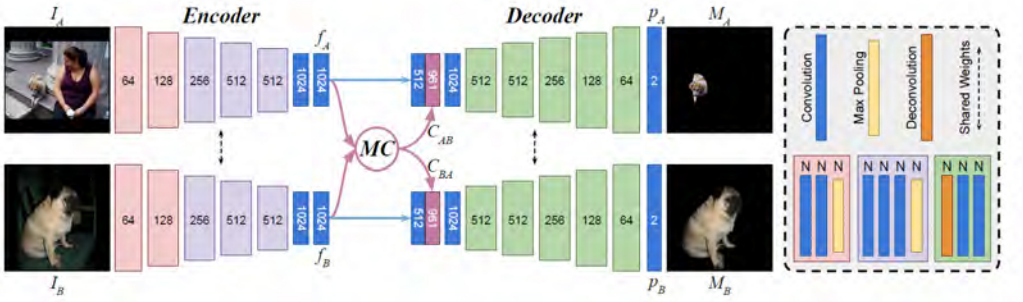

Figure 2.12: The illustration of a deep CNNs for image object co-segmentation (Li, Jafari & Rother 2018).

tion network to obtain correspondence maps, and then, the common objects masks are obtained by passing concatenation of squeezed feature maps and correspondence maps through a Siamese decoder.

2.3 Summary

This chapter reviews two related works: image co-saliency detection and image co-segmentation. In image co-saliency detection review, basic concepts are first introduced, then the basic and closely related work, image saliency detection is presented, and then the works of image co-saliency detection are reviewed. In image co-segmentation review, basic concepts are first introduced, then the close related work, image object segmentation is presented, and then the works of image object co-segmentation is reviewed. In the following chapters, our works and the contributions are explained in details corresponding to which are stated in Introduction section.

Chapter 3

Co-saliency Detection Based on Region-level Fusion and Pixel-level Refinement

3.1 Motivations

The past twenty years have witnessed the booming research on saliency detection and a number of saliency models have been proposed and evaluated (Borji et al. 2015). The object saliency map as well as the object segmentation results serve as an indispensable cue for a variety of saliency-based applications such as human fixationprediction (Itti et al. 1998, Borji 2012, Borji & Itti 2012, Liu, Han, Zhang, Wen & Liu 2015), salient object detection (Borji et al. 2015), salient object segmentation (Liu et al. 2012, Kruthiventi et al. 2016, Li et al. 2017), content-aware image/video retargeting (Shamir & Sorkine 2009, Liang et al. 2016, Cho et al. 2017, Bansal et al. 2018), and content-based image compression (Guo & Zhang 2009, Li, Zuo, Gu, Zhao & Zhang 2018), etc. Most saliency models are designed to detect

salient objects in a single image, but it is still difficult for the state-of-the-art single-image saliency models to effectively highlight salient objects in an image with a complex scene. An interesting issue in the research of object saliency detection is the common objects detection in a group of images. Image co-saliency detection models exploit the repeated occurrence of salient objects in a set of images to better highlight such co-salient objects and suppress irrelevant background regions than single-image saliency detection models. Co-saliency model generates a set of co-saliency maps, which can be exploited for a number of related applications including object co-segmentation (Lee et al. 2015, Tsai et al. 2018), object co-recognition (Cho, Shin & Lee 2008), co-localization (Wei, Zhang, Wu, Shen & Zhou 2019), image matching (Toshev, Shi & Daniilidis 2007, Vo et al. 2019) and image collage (Zhang & Huang 2012).

Some image co-saliency models are proposed to detect co-salient objects in a pair of images in the early work. Local structure changes (Zhang, Han, Li & Wang 2015), weighted Gist (Rubinstein et al. 2013), sift flow (Rubinstein et al. 2013), object frequency (Zhu et al. 2013), etc., are captured for global consistency. However, these models are limited to image pairs. For an image set containing multiple images, the common salient objects should appear frequently in a image set, i.e. appear in most of the images, the measurement of the frequency should be more general and the computation cost should also be considered.

Thus, the image co-saliency model we propose in this chapter is calculated based on regions, for the reasons that regions can provide more local information than pixels and the amount of regions is much smaller than that of pixels, thus, the region-based calculation can reduce the computation cost.

For a region in an image, we define its consistency in another image as its

similarity to its most similar region in this image, and its global consistency will be a fusion of the consistencies in all other images. For a region, the higher of the global consistency value, the more frequently it appears in other images. After the single image saliency map and the global consistency cue are obtained, fusing the global consistency (inter-image saliency) and the single image object saliency (intra-image saliency) into co-saliency is the key step. Some of the image co-saliency models sum up or weighted sum up the two kinds of saliency maps together (Tan et al. 2013, Rubinstein et al. 2013), some of the models multiply them (Liu et al. 2013), and we fuse these two cues by considering not only the co-saliency value of the region own but also the co-saliency value of all its similar regions across the whole image set. The final co-saliency results are obtained by refining the coarse region-level results into pixel-level results, which further improves the performance.

There are two main contributions of the proposed model:

- An image co-saliency model by combining region-level fusion and pixel-level refinement is proposed and the proposed model achieves a better co-saliency detection performance on two benchmark datasets compared to the state-of-the-art co-saliency models.

- A novel measurement and fusion scheme is proposed, which effectively exploits similarities between regions and global consistency over the image set, to fuse intra-image and inter-image cues over the image set into co-saliency; and exploits the similarities between pixel and region to refine the co-saliency cues into final co-saliency results.

The rest of this chapter is organized as follows: Section 3.2 describes the proposed method in detail. Section 3.3 presents the experimental results and analysis. Finally, Section 3.4 gives the conclusion.

3.2 Co-saliency Detection Model

The proposed co-saliency model is illustrated using the co-saliency detection on the example image set shown in Figure 3.1. The following three subsections will describe the generation process of intra saliency maps, region-level co-saliency maps, and pixel-level co-saliency maps in turn.

3.2.1 Intra saliency

In natural images, salient object pixels usually distribute more sparsely than background pixels, which usually have a wider spatial distribution. Therefore, in terms of pixel features such as color and orientation, background pixels usually show a higher correlation than object pixels, and such a difference in the feature space can be effectively exploited to identify salient regions in the image (Shen & Wu 2012). Note that any single-image saliency model can be used for saliency detection in each individual image. In this chapter, we use the segmentation driven low-rank matrix recovery (SLR) model (Zou et al. 2013) to measure the intra saliency for each image in the image set.

Give the image set $\{I_m\}_{m=1}^{M}$, the SLR model is first used to generate for each image the raw saliency map. Then for each image I_m, the gPb-owt-ucm method (Arbelaez et al. 2011) which exploits the globalized probability of boundary based contour detector and the oriented watershed transform to generate the real-valued ultrametric contour map U_m , in which a higher value of a boundary segment indicates a higher boundary strength. By thresholding U_m, a set of boundaries are retained to form a region segmentation result. Specifically, for each normalized into the range of [0, 1], the threshold is gradually increased from 0.1 until the number of the generated regions is



Figure 3.1: Illustration of the proposed co-saliency model. From top to bottom: the image set (5 out of 36 images are shown), ultrametric contour maps, region segmentation results, intra saliency maps, initial region-level co-saliency maps, final region-level co-saliency maps, and pixel-level co-saliency maps.

less than 200, to obtain an over-segmentation result for most natural images. Then for each segmented region $r_{m,i}(i = 1, ..., N_m)$ in each image I_m, its intra saliency measure is calculated as follows:

$$S^{int}(r_{m,i}) = \frac{1}{|r_{m,i}|} \sum_{p \in r_{m,i}} S_m^{raw}(p) \tag{3.1}$$

where p denotes each pixel in $r_{m,i}$, and $|r_{m,i}|$ denotes the number of pixels in $r_{m,i}$.

In Figure 3.1, for the example image set shown in the 1st row, the ultrametric contour maps are shown in the 2nd row, and the region segmentation results are shown in the 3rd row, in which each region is represented using its mean color. Note that in this image set containing 36 images, the co-salient objects are the players of red team. The intra saliency maps are shown in the 4th row, in which most regions of different objects including players of both teams and referees are highlighted, while some background regions in some images are not effectively suppressed.

3.2.2 Region-level co-saliency

All images in the image set $\{I_m\}_{m=1}^{M}$ are transformed into the Lab color space, in which the luminance channel and the two chrominance channels are well decorrelated. Based on the colors of all pixels in the image set, each color channel is uniformly quantized into q bins, to obtain a color quantization table Q with q^3 entries. The parameter q is set to a moderate value, 16, which is generally sufficient for color quantization of natural images. The quantized color of each entry in Q is calculated as the mean color of those pixels falling into the entry. By removing those zero-value entries, Q is updated to have a total of K (usually $K \ll q^3$) entries. A mapping function is defined as $k = Q(c_p)$, where c_p denotes the color of each pixel p, and k

denotes the entry number in Q.

For each region $r_{m,i}$, its color histogram $H_{m,i}$ is calculated based on the quantized colors of all pixels in $r_{m,i}$, and normalized to make $\sum_{k=1}^{k} H_{m,i}(K) = 1$. Then for each pair of regions, $r_{m,i}$ and $r_{n,j}$, the inter-region dissimilarity is defined as the chi-squared distance between $H_{m,i}$ and $H_{n,j}$, i.e.,

$$\phi_R(r_{m,i}, r_{n,j}) = \frac{1}{2} \cdot \sum_{k=1}^{K} \frac{[H_{m,i}(k) - H_{n,j}(k)]^2}{H_{m,i}(k) + H_{n,j}(k)} \tag{3.2}$$

Since a co-salient region in one image can usually find similar regions in other images of the image set, the inter-region dissimilarity between different regions are exploited to find for each region the most similar region in each of the other images, and then used to measure for each region the global consistency over the image set as follows:

$$C(r_{m,i}) = 1 - \frac{1}{M-1} \sum_{n=1, n \neq m}^{M} \min_{j=1,\ldots,N_n} [\phi_R(r_{m,i}, r_{n,j})] \tag{3.3}$$

Using Equation 3.3, the global consistency measure for each region $r_{m,i}$ falls into the normalized range of [0, 1], and is used to multiply with the intra saliency measure to serve as an initial estimate of region-level co-saliency measure, i.e.,

$$S^{co-i}(r_{m,i}) = S^{int}(r_{m,i}) \cdot C(r_{m,i}) \tag{3.4}$$

to suppress the saliency of those regions with a lower global consistency measure. Then, by fully exploiting the similarity between different regions over the image set, the final region-level co-saliency measure for each region $r_{m,i}$ is defined as follows:

$$S^{co-f}(r_{m,i}) = \frac{\sum_{n=1}^{M} \sum_{j=1}^{N_n} [1 - \phi_R(r_{m,i}, r_{n,j})] \cdot S^{co-i}(r_{n,j})}{\sum_{n=1}^{M} \sum_{j=1}^{N_n} [1 - \phi_R(r_{m,i}, r_{n,j})]} \tag{3.5}$$

where the weight $[1 - \phi_R(r_{m,i}, r_{n,j})]$ with the range of $[0, 1]$ actually represents the similarity between $r_{m,i}$ and $r_{n,j}$, accordingly.

On the basis of intra saliency maps shown in the 4th row of Figure 3.1, the initial and final region-level co-saliency maps are shown in the 5th and 6th row, respectively. We can observe that the co-salient objects, i.e., the players of red team, are highlighted, while irrelevant regions including the players of white team, referees, and background regions such as playfield and trees are suppressed in the final region-level co-saliency maps.

3.2.3 Pixel-level co-saliency

Based on final region-level co-saliency measures of each image, the pixel-level saliency map is derived by exploiting the color-spatial similarity between each pixel and its neighboring regions, and regulated by the object prior map, which is evaluated based on region's connectivity with image borders. Specifically, for each pixel $p \in r_{m,i}$, its co-saliency is defined as the weighted sum of co-saliency measures of its neighboring regions with the regulation factor of object prior as follows:

$$S_m^{co-p}(p) = O(p)\frac{\sum_{r_{m,j} \in N_p} \Psi(p, r_{m,j}) \cdot S^{co-f}(r_{m,j})}{\sum_{r_{m,j} \in N_p} \Psi(p, r_{m,j})} \tag{3.6}$$

where $O(p)$ denotes the object prior for pixel p, N_p denotes the region-level neighborhood for the pixel$p$, and $\Psi(p, r_{m,j})$ denotes the color-spatial similarity between the pixel p and the region $r_{m,j}$. Specifically, N_p includes $r_{m,j}$ which contains p and those regions adjacent with $r_{m,j}$.

The color-spatial similarity $\Psi(p, r_{m,j})$ is defined as follows:

$$\Psi(p, r_{m,j}) = \begin{cases} H_{m,j}[Q(c_p)], & j = i \\ H_{m,j}[Q(c_p)] \cdot exp\left(-\frac{\|x_p - y_{m,j}\|_2}{\|x_p - y_{m,i}\|_2}\right), & j \neq i \end{cases} \tag{3.7}$$

where x_p denotes the spatial position of pixel p, and $y_{m,j}$ denotes the spatial center position of region $r_{m,j}$. Equation 3.7 indicates that a higher weight is assigned to the region, which shows a higher color-similarity with the pixel p, i.e., a higher occurrence of the quantized color of pixel p in the region's color histogram, and also locates spatially nearer to the pixel p. Therefore, the neighboring regions contribute to the saliency calculation of the pixel p according to their color-spatial similarities with the pixel p.

The object prior $O(p)$ is defined based on the observation that background regions usually connect with image borders in natural images, while salient object regions do not touch image borders or connect with image borders less than background regions. As suggested in (Zou et al. 2013), to calculate the object prior based on this observation, a relatively coarse region segmentation, which partitions homogenous background into fewer regions, is suitable to suppress those background regions connecting with image borders more uniformly. Therefore, the ultrametric contour map U_m for each image I_m is thresholded to obtain a coarse region segmentation result with just less than 20 regions, and the object prior for each coarsely segmented region $R_{m,j}$ is defined as follows:

$$O(R_{m,i}) = 1 - \left[\frac{R_{m,i} \cap B_m}{prm(R_{m,i})}\right]^2 \tag{3.8}$$

where $prm(R_{m,i})$ denotes the outer perimeter of $R_{m,i}$, and B_m denotes the borders of image I_m. For each pixel p, its object prior is assigned with that of the region containing p, i.e., $O(p) = O(R_{m,i}), \forall p \in R_{m,i}$.

On the basis of final region-level co-saliency maps in the 6th row of Figure 3.1, the generated pixel-level co-saliency maps shown in the bottom row of Figure 3.1 further suppress irrelevant regions and preserve well-defined boundaries of co-salient object regions.

3.3 Experiments

We performed experiments on two public datasets with binary ground truths for co-salient objects. Specifically, the Co-saliency Pairs (CP) dataset (Li & Ngan 2011) contains 210 images, i.e.,105 image pairs, and the CMU Cornell iCoseg dataset (Batra et al. 2010) contains 643 images from 38 object classes, each of which has 5 to 41 images. We compared the co-saliency detection performance with three state-of-the-art co-saliency models, i.e., Li's model (Li & Ngan 2011), Fu's model (Batra et al. 2010) and Liu's model (Liu et al. 2013). Note that Li's model is only suitable for co-saliency detection on image pairs, and thus only tested on the CP dataset.

For subjective comparison, co-saliency maps generated for some image pairs in the CP dataset and some image sets in the iCoseg dataset are shown in Figure 3.2, Figure 3.3 and Figure 3.4, Figure 3.5, Figure 3.6, Figure 3.7, respectively. For a fair comparison, all co-saliency maps are normalized into the same range of [0, 255]. For some image pairs and image sets with relatively homogenous background, such as the three image pairs in Figure 3.2 and the image set in Figure 3.4, co-salient objects are correctly highlighted using most models, while our model can highlight the complete co-salient objects more uniformly compared to other models. For the other image pairs/sets in Figures, which contain non-homogeneous objects (leopards in Figure 3.3), large-scale objects (stonehenges in Figure 3.6), complex background (the image pairs in Figure 3.3 and the image sets in the Figure 3.5 and Figure 3.7), our model highlights the complete co-salient objects and suppresses irrelevant regions more effectively compared to other co-saliency models.

For objective comparison, we performed thresholding with a series of fixed integers from 0 to 255 on co-saliency maps to obtain a set of binary co-

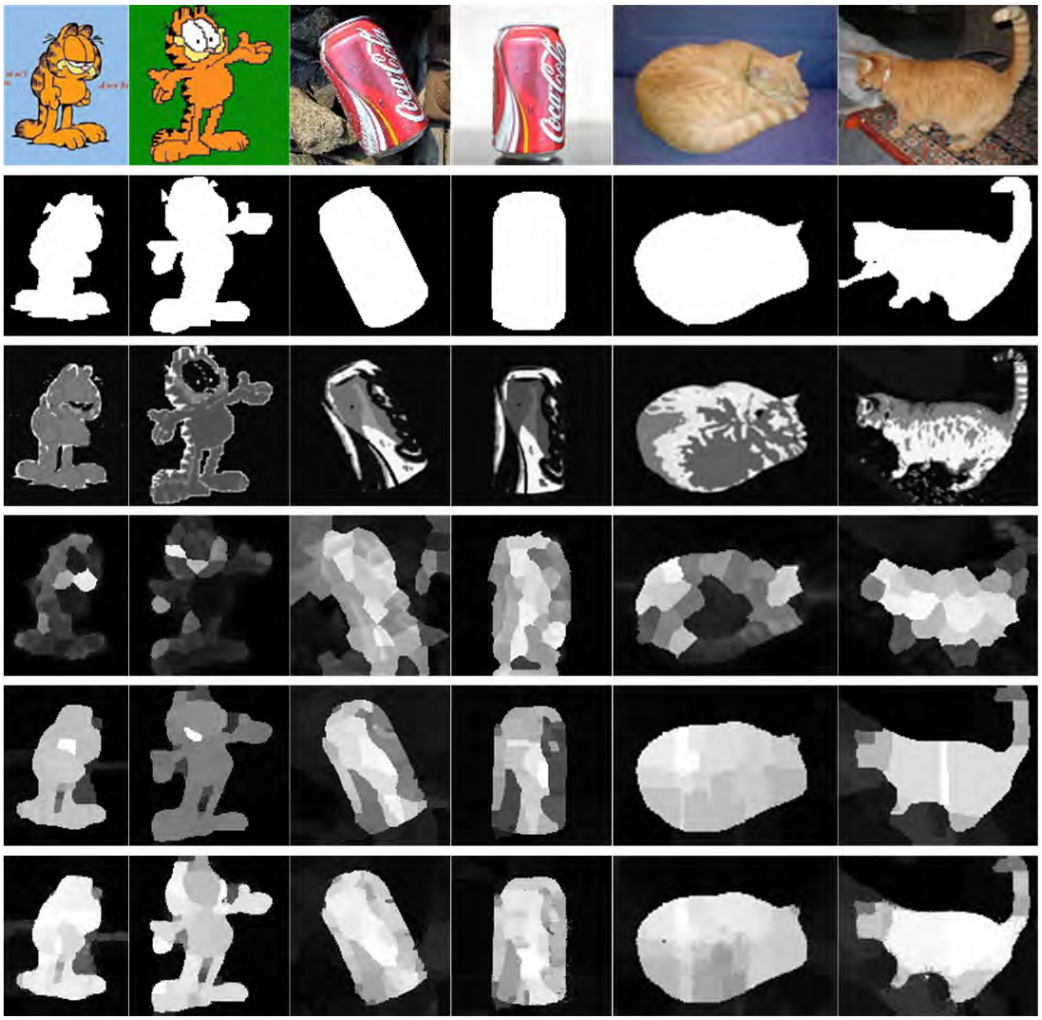

Figure 3.2: Co-saliency maps generated for image pairs in the CP dataset. From top to bottom: original image pairs, binary ground truths, co-saliency maps generated using Li's model (Li & Ngan 2011), Fu's model (Batra et al. 2010), Liu's model (Liu et al. 2013) and our model, respectively.

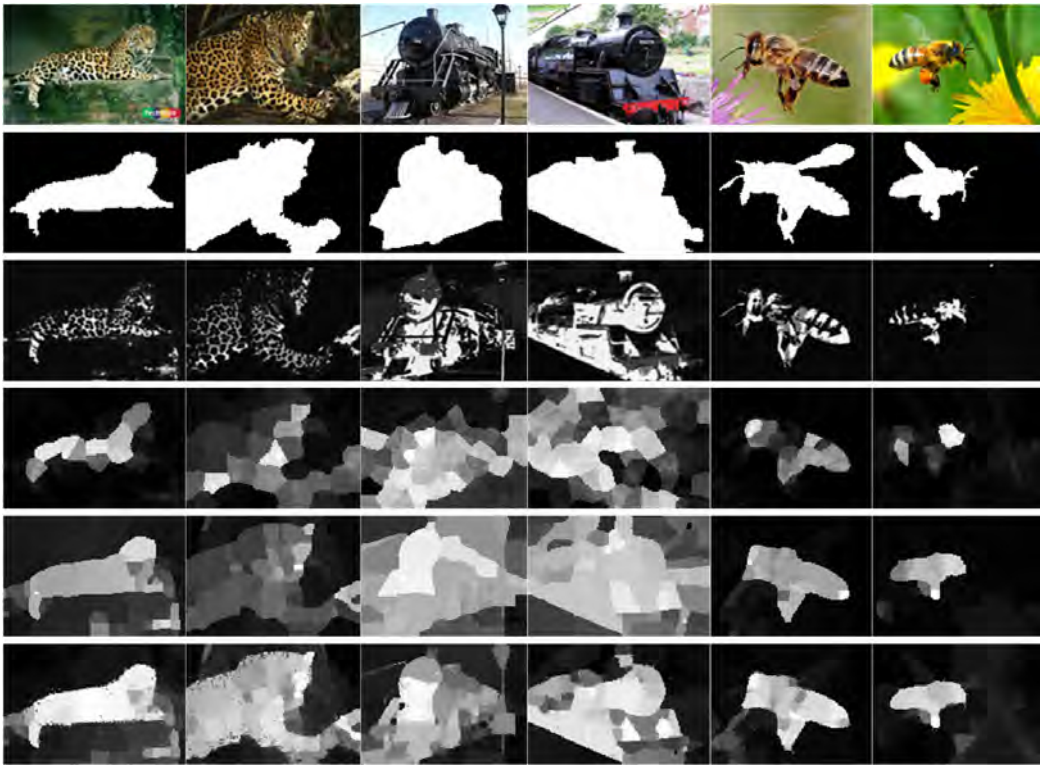

Figure 3.3: Co-saliency maps generated for image pairs in the CP dataset. From top to bottom: original image pairs, binary ground truths, co-saliency maps generated using Li's model (Li & Ngan 2011), Fu's model (Batra et al. 2010), Liu's model (Liu et al. 2013) and our model, respectively.

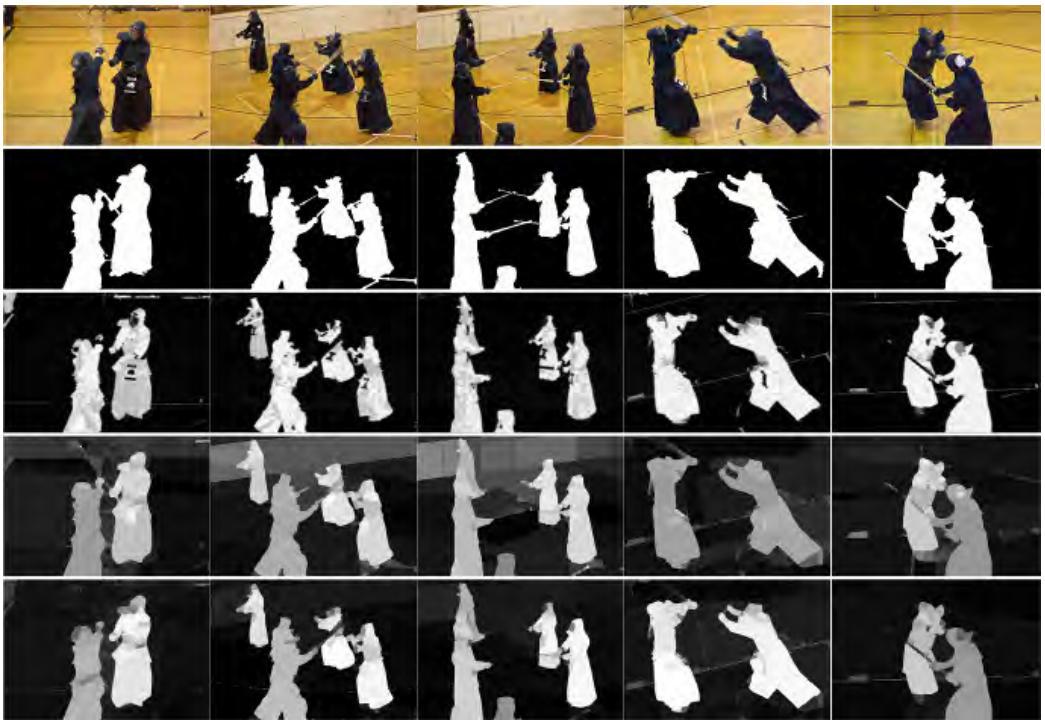

Figure 3.4: Co-saliency maps generated for an image set in the iCoseg dataset. From top to bottom in each sub-figure: some original images in the image set, binary ground truths, co-saliency maps generated using Fu's model (Batra et al. 2010), Liu's model (Liu et al. 2013) and our model, respectively.

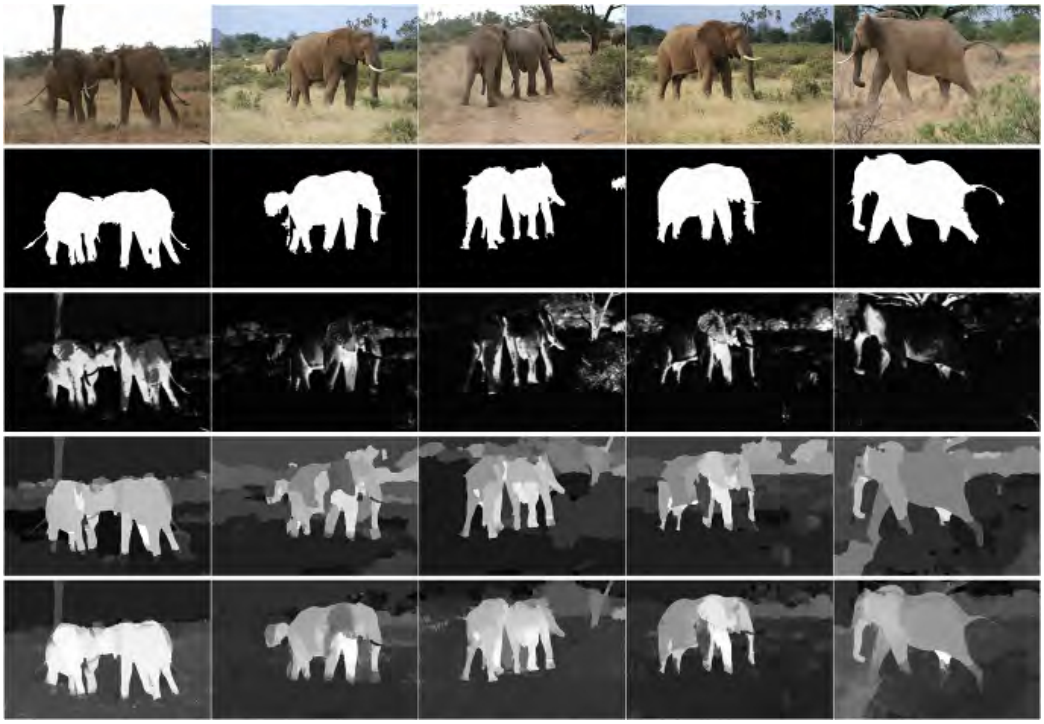

Figure 3.5: Co-saliency maps generated for an image set in the iCoseg dataset. From top to bottom in each sub-figure: some original images in the image set, binary ground truths, co-saliency maps generated using Fu's model (Batra et al. 2010), Liu's model (Liu et al. 2013) and our model, respectively.



Figure 3.6: Co-saliency maps generated for an image set in the iCoseg dataset. From top to bottom in each sub-figure: some original images in the image set, binary ground truths, co-saliency maps generated using Fu's model (Batra et al. 2010), Liu's model (Liu et al. 2013) and our model, respectively.



Figure 3.7: Co-saliency maps generated for an image set in the iCoseg dataset. From top to bottom in each sub-figure: some original images in the image set, binary ground truths, co-saliency maps generated using Fu's model (Batra et al. 2010), Liu's model (Liu et al. 2013) and our model, respectively.

salient object masks, and calculated the precision and recall measures using the binary ground truths in the two datasets as reference masks. Then for each co-saliency model, at each threshold, the precision/recall values for all co-saliency maps in each dataset are averaged to plot the precision-recall curves.

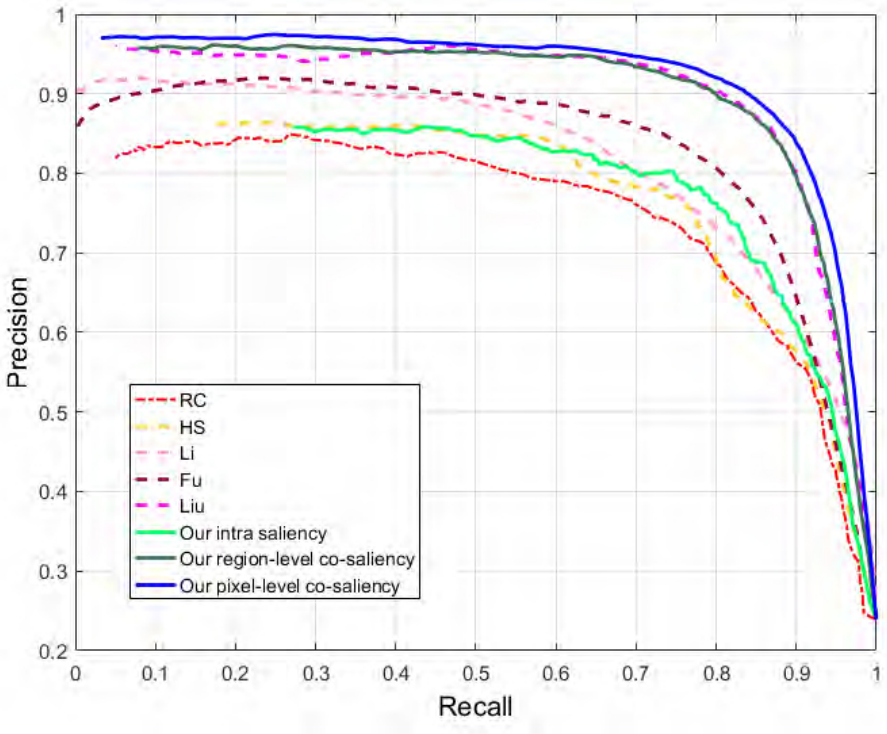

Figure 3.8: Precision-recall curves for different saliency maps on the CP dataset.

As shown in Figure 3.8 and Figure 3.9, the precision-recall curves of our intra, final region-level and pixel-level co-saliency maps show the progressive improvement on co-saliency detection performance, which demonstrates the contribution of region-level fusion and pixel-level refinement, respectively.

Figure 3.8 and Figure 3.9 show that on both CP and iCoseg datasets, the precision-recall curves of our co-saliency maps are consistently higher

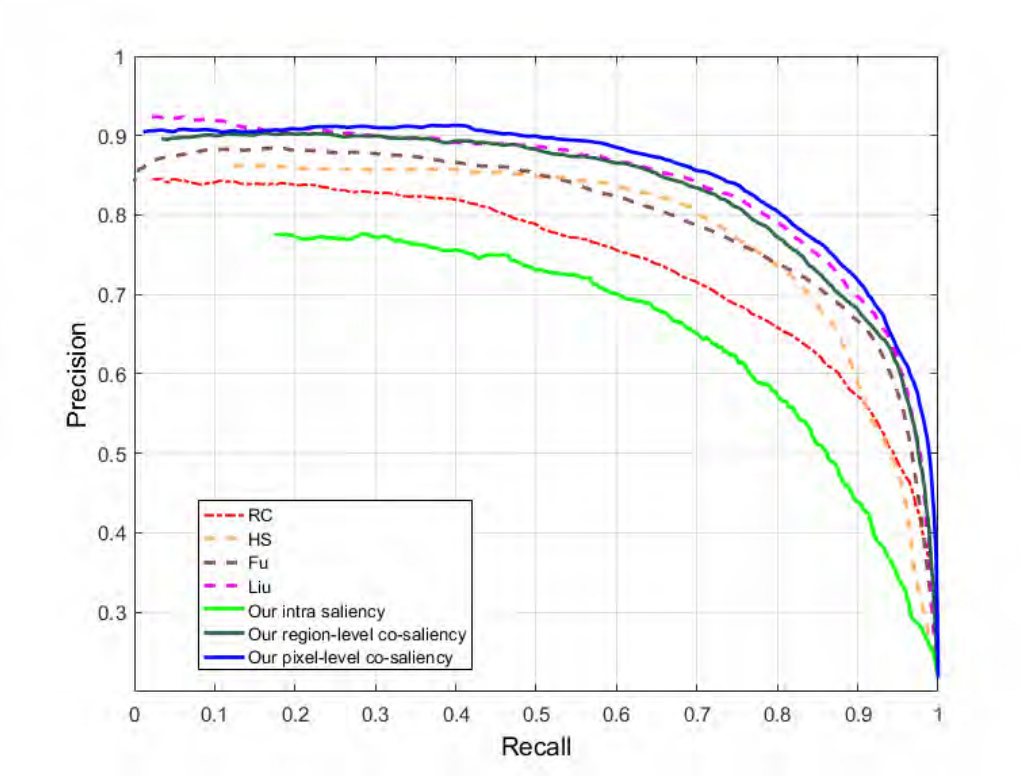

Figure 3.9: Precision-recall curves for different saliency maps on the iCoseg dataset.

than other co-saliency maps, respectively, and thus can objectively demonstrate that our model outperforms other models on the co-saliency detection performance.

Besides, as shown in Figure 3.8 and Figure 3.9, the PR curves of saliency maps generated using two state-of-the-art single-image saliency models, i.e., RC: region contrast (Cheng, Mitra, Huang, Torr & Hu 2014) and HS: hierarchical saliency model (Yan et al. 2013), are lower than those PR curves for co-saliency models, and this clearly shows the advantage of co-saliency detection in a set of images.

Our model is implemented using Maltab on Ubuntu 12.04 with 3.4GHz CPU. The average processing time per image is 16.48s on CP and 109.27s on iCoseg, respectively, and gPb-owt-ucm occupies 87.3% and 83.8% processing time, respectively.

3.4 Summary

This chapter proposes an effective co-saliency model for co-saliency detection in a set of images. In our model, the region-level fusion exploits the similarities between different regions and the global consistency measures over the image set, and the pixel-level refinement utilizes the similarities between region and pixel as well as object priors. Experimental results on two benchmark datasets show the better co-saliency detection performance of our model.

Chapter 4

Unsupervised Image Co-segmentation via Guidance of Simple Images

4.1 Motivations and Significance

Image co-segmentation has received a boosting attention by researchers who work on object segmentation. Object segmentation is a fundamental problem in computer vision for many applications such as content-based image retrieval (Liu & Yang 2013, Gao, Wang, Tao, Ji & Dai 2012), object-aware image/video retargeting (Lin, Yeh, Lin & Lee 2013, Setlur, Lechner, Nienhaus & Gooch 2007), image editing (Zhang, Huang, Li & Metaxas 2012, Liu, Wang, Li, Li, Ma, Lu & Ma 2007) and content-aware image/video compression (Xue, Li & Zheng 2011, Shen, Liu & Zhang 2013). Object segmentation in a single image usually incorporates different measures at the level of pixels (Shi & Malik 2000, Boykov & Jolly 2001, Rother et al. 2004), or at the level of regions (Fulkerson et al. 2009) into a global objective function; by

minimizing the objective function, the objects can be separated from the background. This kind of methods segments the object in each image in isolation, while co-segmentation aims to segment objects with additional information from other images containing the same or similar objects(Rother et al. 2006). Thus, image co-segmentation can be derived as a task of jointly segmenting the common objects in a group of images.

Most of the co-segmentation methods are unsupervised and have various problem formulations(Li et al. 2016, Lin et al. 2015, Joulin et al. 2010, Rubinstein et al. 2013, Chen et al. 2015, Lee et al. 2015, Yu & Qi 2014, Faktor & Irani 2013, Meng et al. 2013, Zou et al. 2015). In (Joulin et al. 2010), the segmentation problem is handled in a discriminative clustering framework, in which objects are separated from the background by merging the image pixels into two clusters that can be maximally distinguished. Disturbance from consistent background is considered in (Li et al. 2016), and a framework of saliency detection and discriminative learning is designed to overcome the disturbance. In (Lin et al. 2015), each image is treated as a player, for which a constrained utility function is designed by using the cooperative game, heat diffusion, and image saliency. Common objects can be segmented according to the labels defined via the constrained utility function. In (Rubinstein et al. 2013), pixel-level correspondences between images are considered, while in (Faktor & Irani 2013, Yu & Qi 2014), it is region-level affinities that are calculated. In (Faktor & Irani 2013), the affinities are calculated to score the multiple overlapping segment candidates, then these candidates and their scores are used to generate co-segmentation likelihood maps. In (Yu & Qi 2014), region-level affinities are utilized for each superpixel to find its correspondingly matched superpixels. Some co-segmentation methods utilize correspondences among all the images (Lee et al. 2015) while

others compute correspondences among specific selected images (Rubinstein et al. 2013). In (Chen et al. 2015), a feature pool is built based on overlapped rectangle regions and grows with the number of processed images, and the image to be processed obtains its prior information via the feature pool. In (Meng et al. 2013), a feature adaptive co-segmentation method is proposed to replace the fixed region similarity measurement for the better region matching. This method first evaluates the image complexities, then selects simple images and obtains the initial segments. This method uses a linear combination of the common features, and the linear combination parameters are learned from these segments. The learnt adaptive feature is used to segment all the images in an image group. In (Zou et al. 2015), a saliency boosting model is proposed to optimize the joint segmentation results and saliency maps mutually.

There are some limitations of the image co-segmentation methods we have mentioned above. For the methods which utilize the correspondences among all the images, they treat the detected/segmented objects in all the images as the same; while our observation indicates that the object detection/segmentation results in complicated images are often unsatisfactory, thus unreliable. For the methods that select similar images first, then calculate the correspondences among the selected images, they may achieve satisfactory results for simple images; but not for complicated images. For it could be quite difficult for these images with clutter background to distinguish objects from background, and the correspondence calculation may be useless, even noisy. These limitations have been noticed in (Chen et al. 2015) and the propagation method from simple images to complicated images has been used, however failure results may also be propagated. In (Meng et al. 2013), images are also first ranked by their complexities, and the main difference

between (Meng et al. 2013) and the proposed method is the use of simple images. (Meng et al. 2013) utilizes simple images to select the features and to replace the fixed region similarity measurement for the better region matching, while the proposed method utilizes simple images to extract samples and guide the complicated image segmentation. Besides, the image ranking method is also different from that in (Meng et al. 2013).

The proposed method pays different regards on simple and complicated images. The simple images are to have better segmentation results which deserve more confidence; while the complicated images, on the contrary, need the aid of simple images. In the proposed method, the complicated images are segmented with the guidance of simple images.

There are two main contributions of the proposed method:

- A new effective method to rank image complexities via their saliency maps.
- An effective framework to utilize simple images to segment complicated images.

The rest of this chapter is organized as follows: Section 4.2 describes the proposed method in detail. Section 4.3 presents the experimental results and analysis. Finally, Section 4.4 gives the conclusion.

4.2 The Proposed Method

The proposed method includes four steps: *Image Ranking, Simple Image Selection, Samples Extraction* and *Complicated Image Segmentation*, as shown in Figure 4.1.

Given an image set$\{I^1, I^2, I^3, ..., I^N\}$ composed of N images in total. Firstly, we obtain their single saliency maps $\{S^1, S^2, S^3, ..., S^N\}$ by a single

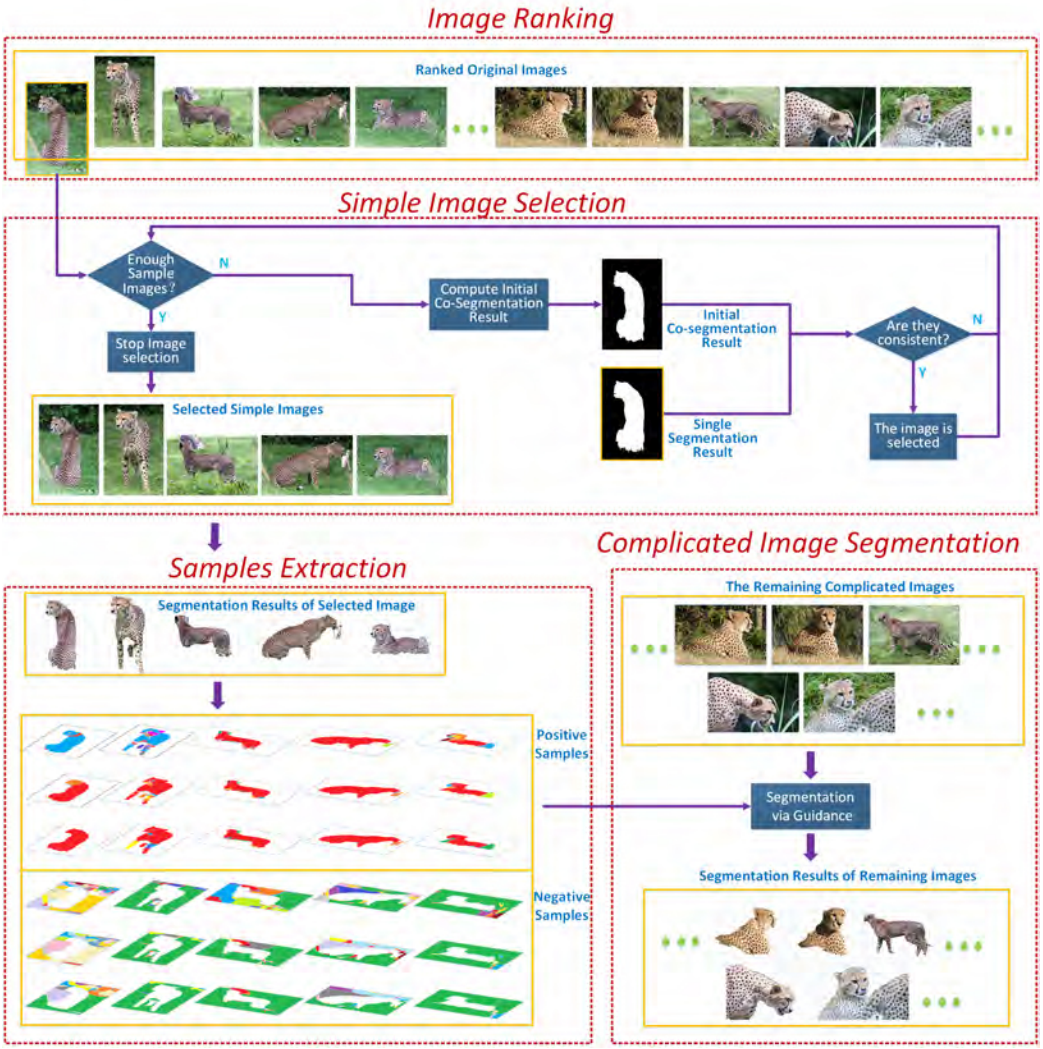

Figure 4.1: The flowchart. There are four steps in the proposed method: image ranking, simple image selection, samples extraction and complicated image segmentation.

saliency detection model (Liu et al. 2014) and rank the images based on these saliency maps. Secondly, we select the images, in which objects are common and easy to be segmented, as simple images. Thirdly, we co-segment the simple images to obtain their segmentation results and extract the result pieces as samples both positive and negative. Finally, the remaining complicated images are segmented with the guidance of the obtained samples.

To illustrate the whole framework clearly, we introduce some pre-processing work first.

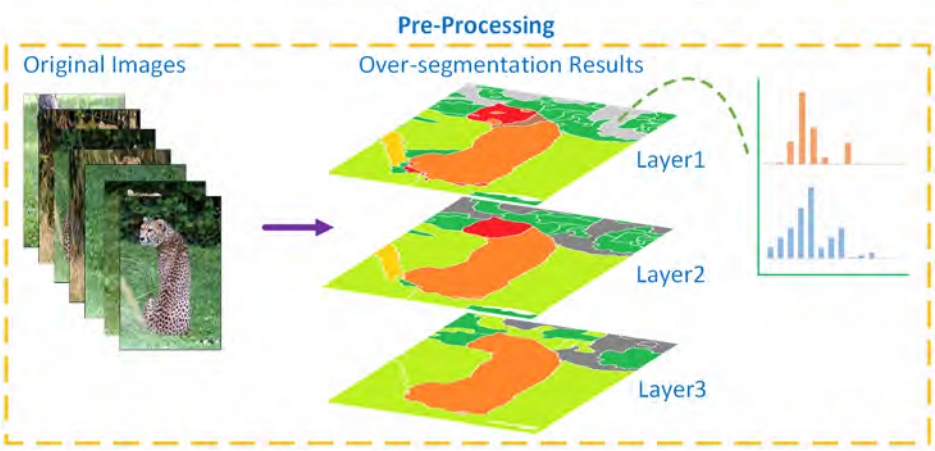

Figure 4.2: The pre-processing. One image is over-segmented into hierarchical regions in three layers and is represented by two histograms: color histogram and dense SIFT histogram.

Pre-processing For each pixel in an image group, we extract its color feature in both RGB and Lab color space, totally in 6 dimensions; and every 4 pixels we extract a dense SIFT descriptor (Vedaldi & Fulkerson 2010) in 128 dimensions. We cluster all the color features into K_c clusters, and all the dense SIFT features into K_s clusters using k-means clustering (Lloyd 1982). Here, K_c and K_s are determined by the two minimum distances, which are set

to moderate values, 0.15 and 0.7, for the normalized colour features and the normalized SIFT features, respectively. Specifically, during the clustering process, the distance between each cluster pairs is computed, and if there exists a distance shorter than the minimum distance, the closest cluster pair will be merged until no distance between cluster pairs is shorter than the minimum distance. Each pixel has a color cluster label and a dense SIFT cluster label. As shown in Figure 4.2, we over-segment each image into three-layer hierarchical regions using gpb-owt-ucm method (Arbelaez et al. 2011). For each image, in layer 1, there will be most regions; in layer 2, nearly a half; in layer 3, there will be only about a quarter regions. For an image, there are M regions at most in layer 1 and for a region in each layer, it can be represented by two histograms: color histogram H_c with K_c bins and dense SIFT histogram H_s with K_s bins.

For region $R^{i,l,n}$, which denotes the ith region in the lth layer of the nth image, its color histogram is denoted as $H_c^{i,l,n}$ and its dense SIFT histogram as $H_s^{i,l,n}$. For two arbitrary regions $R^{i,l,n}$ and $R^{j,k,m}$ in two different images, as shown in Figure 4.3, their dissimilarity is calculated as the chi-square distance between their feature histograms as follows:

$$D(R^{i,l,n}, R^{j,k,m}) = 0.5 \cdot (\chi^2(H_c^{i,l,n}, H_c^{j,k,m}) + \chi^2(H_s^{i,l,n}, H_s^{j,k,m})) \tag{4.1}$$

These colour and SIFT features are used for the purpose of complementing each other. Figure 4.4 shows some experimental region matching results. As we can see, regions in different layers can successfully find their most similarity regions.

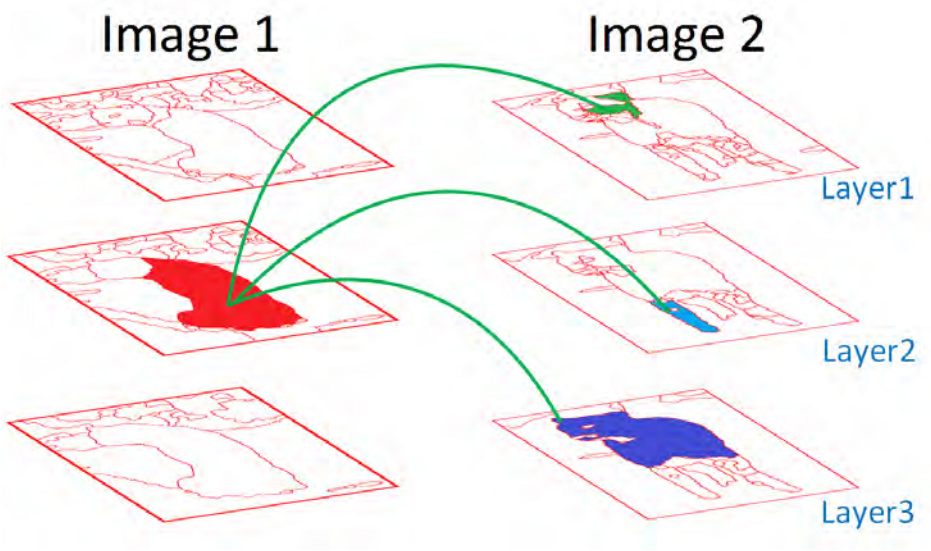

Figure 4.3: Region matching. For one region in layer 2 of image 1, its most similar region is searched through all the three layers in image 2.

4.2.1 Image ranking

The goal of image ranking is to find the images in which objects are easy to be segmented. We rank images inspired by (Mai & Liu 2014) based on saliency maps. As shown in the right part of Figure 4.5, the top ranked images (e.g. the five images in the 1st row) have distinct and uniform saliency maps while the lower ranked ones (e.g. the 2 images in the 3rd row) have bad saliency maps and the objects are not easy to be separated.

The ranking score consists of four elements: *Coverage Score, Compactness Score, Distribution Score* and *Contrast Score*.

Given a single saliency map S, it is separated into foreground S_f and background S_b by OTSU method (Otsu 1975), i.e.,$\{S_f, S_b\} = OTSU(S)$.

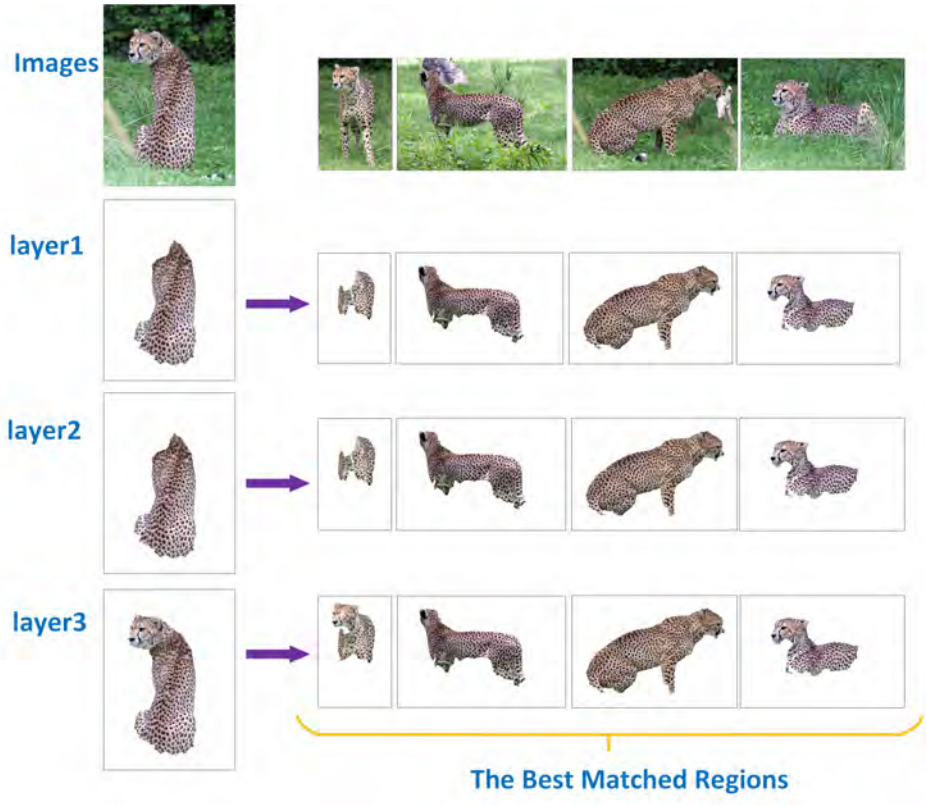

Figure 4.4: The region matching results. The 1st row shows original images, the 1st column shows the regions hierarchically extracted from the three layers, and the 2nd-5th column show the corresponding best matched regions in other images.

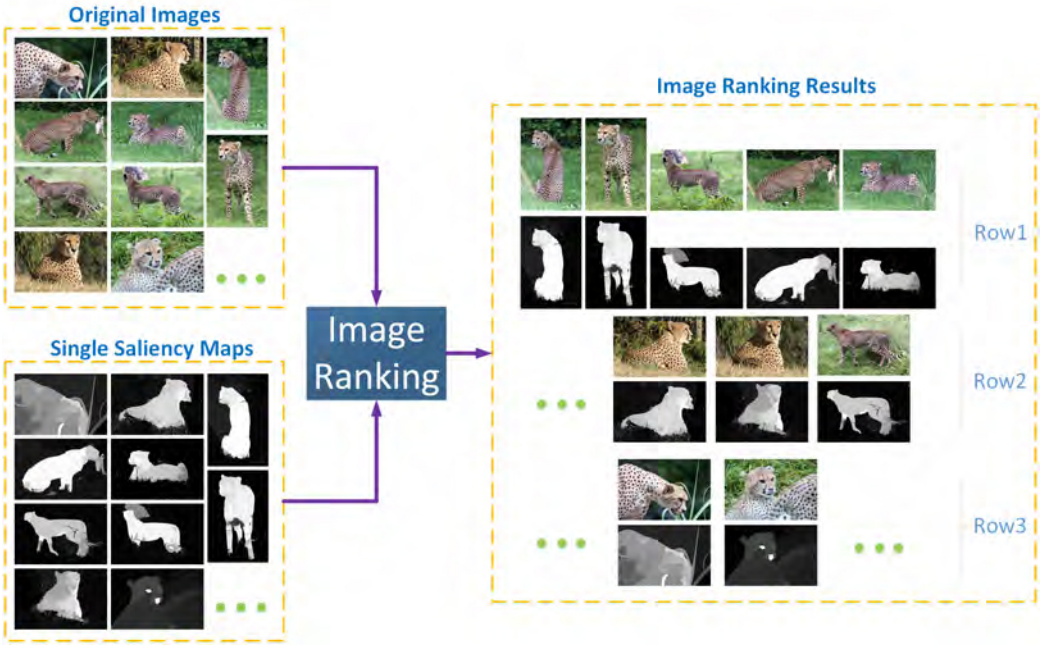

Figure 4.5: Image Ranking. Images are ranked based on their saliency maps.

Coverage Score

In a high-quality saliency map, the salient region should not cover too much or too little of the whole area. According to the statistical results in (Mai & Liu 2014), we find that the percentage of the S_f occupying the whole area may follow a Gaussian distribution, with the mean value around 0.2 and variance value around 0.2. $P_{cov} = \frac{|S_f|}{|S|}$ and r_{cov} are denoted as the percentage, and coverage score respectively, then

$$r_{cov} = -\frac{(P_{cov} - \mu)^2}{\sigma^2} \tag{4.2}$$

where $\mu = 0.2$,$\sigma = 0.2$, they are set based on the statistical results in (Mai & Liu 2014).

Compactness Score

A good saliency map should concentrate its salient pixels in a compact region. We search for the corresponding minimum window w_i when the sum of saliency values in w_i occupies a specific percentage of the sum of saliency values in S (denote the percentage as P^i_{com}). Then we compute the average saliency value of w_i (denoted as S^i_{com}). Three windows are searched with $P^i_{com} \in \{0.25, 0.5, 0.75\}, i = 1, 2, 3$. The compactness score r_{com} is computed as a weighted summation as

$$r_{com} = \frac{\sum_i P^i_{com} \cdot S^i_{com}}{\sum_i P^i_{com}} \tag{4.3}$$

Distribution Score

A good saliency map should have two distinct peaks of its value statistics. One for the background regions is close to 0 and the other for the foreground regions is close to 1. To measure the distribution of a saliency map, we first compute two histograms, H_f and H_b for S_f and S_b, respectively. We set the two histograms both with B_d bins. For the bth bin with the histogram value h_b, its weighted histogram value $r_f(b)/r_b(b)$ for foreground/background is calculated by

$$\begin{cases} r_f(b) = h_b \cdot e^{0.5 \cdot (b - B_d)}, b \in H_f \\ r_b(b) = h_b \cdot e^{-0.5 \cdot (b-1)}, b \in H_b \end{cases} \tag{4.4}$$

where $b = 1, 2, ..., B_d$. In our experiments, we set $B_d = 8$. Equation 4.4 indicates that in the foreground histogram, one bin which is closer to bin B_d has a higher weight; in the background histogram, one bin which is closer to the 1st bin has a higher weight. The distribution score is calculated as

$$r_{dis} = \sum_{b=1}^{B_d} r_f(b) + \sum_{b=1}^{B_d} r_b(b) \tag{4.5}$$

A saliency map with distinct peaks in the value statistics can achieve a high distribution score.

Contrast Score

A simple image should have a large contrast between its foreground and background regions. For each image, RGB channels are quantized into ${B_c}^3$ bins, with each channel quantized into B_c bins. Then the foreground region S_f and background region S_b are transformed into normalized color histograms hc_f and hc_b, respectively. The contrast score is computed as the chi-square distance between the two histograms

$$r_{con} = \chi^2(hc_f, hc_b) \tag{4.6}$$

The larger the distance, the higher the contrast score.

Final Ranking Score

Finally, the ranking score is defined as

$$R = r_{cov} \cdot r_{com} \cdot r_{dis} \cdot r_{con} \tag{4.7}$$

for the reason that images which can simultaneously obtain four high scores may have good quality. The higher the ranking score, the better quality the images may have.

Figure 4.6 shows some ranking results based on different saliency maps. The three kinds of saliency maps are obtained using three state-of-the-art saliency models: ST (Liu et al. 2014), HS (Yan et al. 2013) and RBD (Zhu et al. 2014). The experimental results [1] show that the ranking results vary

[1]Full image ranking results: `https://drive.google.com/open?id=0B9WIl2gJ2teEWnZodUMwS1kyQTA`

with the saliency maps, but in fact there is no essential difference. Although some images will change their ranks, simple images will not be ranked to the bottom as very complicated images, and complicated images will not be ranked to the top as simple images to guide the segmentation of the other images. Therefore, our image ranking method is overall robust to saliency maps generated using different saliency models.

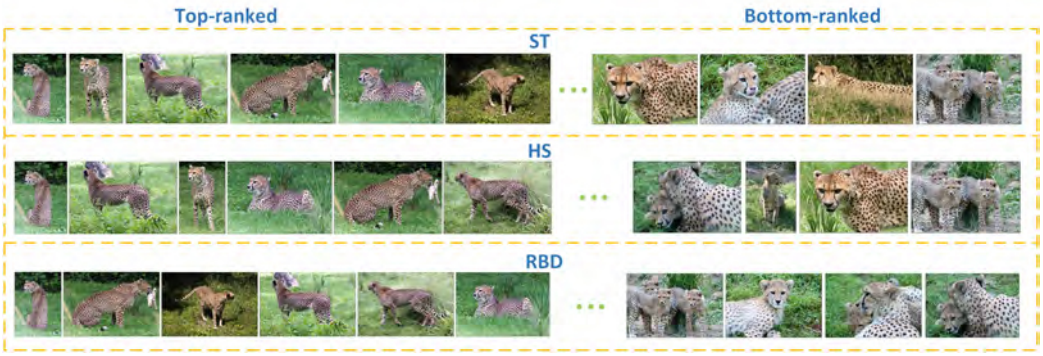

Figure 4.6: Image ranking results with different saliency models. The 1st, 2nd and 3rd row show the ranking results based on saliency maps generated using ST (Liu et al. 2014), HS (Yan et al. 2013) and RBD (Zhu et al. 2014), respectively.

4.2.2 Simple image selection

In this section, simple image selection aims to select a few images that contain common and easily segmented objects, so their segmentation results can guide the segmentation of the others. Although we have already ranked the images, there may be some noisy images in the top ones that highlight the uncommon regions in the image groups. As shown in Figure 4.7, the first ranked 4 images are not selected as simple images for uncommon regions (the red and the blue clouds) are detected.

We use global consistency to guarantee that the segmented objects from

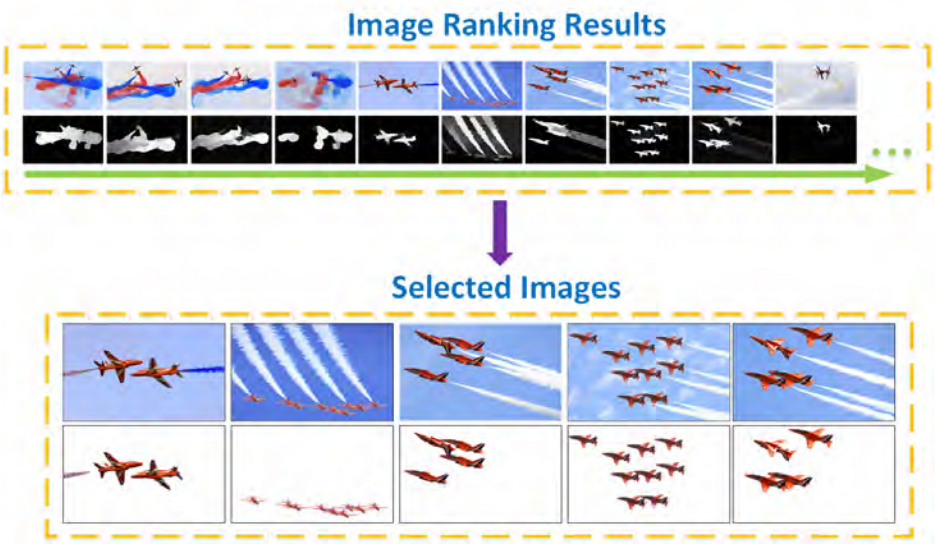

Figure 4.7: Simple image selection. There are two boxes. Top box: the original ranked images and their corresponding single saliency maps. Bottom box: the selected simple images and their corresponding co-segmentation results.

the simple images are common in the whole image set.

Global consistency for each region is calculated among the first half of the ranked images. There are two advantages: 1) to reduce the computation cost and 2) the images among the first half are more reliable compared with those in the second half. For region $R^{i,l,n}$, the consistency value is calculated as

$$C^{i,l,n} = 1 - \frac{2}{N} \sum_{m=1}^{N/2} \min_{j,k} D(R^{i,l,n}, R^{j,k,m}) \tag{4.8}$$

which indicates to find the most similar regions of $R^{i,l,n}$ in other images, then sum up the minimum distances with normalization to obtain the global consistency. A higher global consistency value means a higher frequency that the region appears over the image group. By combining consistency value computed by Equation 4.8 with its single saliency value $S_{si}^{i,l,n}$, we can obtain

an initial object likelihood $P^{i,-l,n}$

$$P^{i,-l,n} = S_{si}^{i,l,n} \cdot C^{i,l,n} \tag{4.9}$$

Since we use hierarchical segmentation method, as Figure 4.2 shows, i.e., a region in layer 3 can be split into several regions in layer 2 and a region in layer 2 can be split into several regions in layer 1. Following the (Yan et al. 2013) the regions in the different layers can be viewed as nodes with parent-child relationship in a tree structure graph. Define $P^{i,l,n}$ as object likelihood variable of node $R^{i,l,n}$, and the set P^n contains all variables in image I^n. Like (Yan et al. 2013), we construct an energy function which consists of two parts: data term $E_D(R^{i,l,n})$ and hierarchy term $E_S(R^{i,l,n}, R^{j,l+1,n})$

$$E(P^n) = \sum_l \sum_i E_D(R^{i,l,n}) + \sum_l \sum_{i,R^{i,l,n} \subseteq R^{j,l+1,n}} E_S(R^{i,l,n}, R^{j,l+1,n}) \tag{4.10}$$

The data term E_D is to gather confidence in each layer, and is defined for every node as follows:

$$E_D(R^{i,l,n}) = ||P^{i,l,n} - P^{i,-l,n}||_2^2 \tag{4.11}$$

where $P^{i,-l,n}$ is the initial object likelihood value of region $R^{i,l,n}$ calculated by Equation 4.9. The hierarchy term E_S considers the consistency between parent and child nodes, and is defined as follows:

$$E_S(R^{i,l,n}, R^{j,l+1,n}) = ||P^{i,l,n} - P^{j,l+1,n})||_2^2 \tag{4.12}$$

By minimizing the energy function Equation 4.10, we can obtain the fusion result of initial object likelihood. Since the energy function is a convex function, it can be solved via belief propagation.

The initial co-segmentation result of image I^n can be obtained by separating object from the background in layer 1 of P^n, which is denoted as $P^{top,n}$

$$F_{rc} = OTSU(P^{top,n}) \tag{4.13}$$

We aim to select N_s simple images by comparing the single segmentation result with the initial co-segmentation result. Denote $F_{si} = OTSU(S_{si})$ as the single segmentation result, then we use their Intersection-over-Union(IoU) as the selection criterion

$$IoU = \frac{|F_{si} \bigcap F_{rc}|}{|F_{si} \bigcup F_{rc}|} \geq \theta \tag{4.14}$$

From the first image in the ranking queue, if its IoU is larger than or equal to θ, the two segmentation results are considered to be similar and reliable, and the image is selected as a simple image. Repeat this step until we obtain N_s simple images. If there are not enough N_s images that can meet Equation 4.14, the images will be selected by their IoU values from top to down, to make the selected simple images as reliable as possible.

4.2.3 Samples extraction

After selecting simple images, we then segment these images and extract samples from the segmentation results.

For each region $R^{i,l,n}$ in the nth simple image, we first search for its most similar regions in the mth simple image as follows:

$$K(m), J(m) = arg \min_{k,j} D(R^{i,l,n}, R^{j,k,m}) \tag{4.15}$$

Thus, the common object likelihood value of region $R^{i,l,n}$ can be calculated as follows:

$$P_{co}^{i,-l,n} = \sum_{m=1}^{N_s} P^{J(m),-K(m),m} \cdot (1 - D(R^{i,l,n}, R^{J(m),K(m),m})) \tag{4.16}$$

where $P^{J(m),-K(m),m}$ is the initial object likelihood calculated by Equation 4.9.

For region $R^{i,l,n}$, we have obtained its common object likelihood value $P_{co}^{i,-l,n}$, then we obtain the optimized common object likelihood $P_{co}^{i,l,n}$ for

region $R^{i,l,n}$ by minimizing the tree-structure energy function Equation 4.10. Denote P_{co}^{n} as the set that contains all the values in image I^n, and $P_{co}^{top,n}$ as the values in layer 1.

We follow the framework of Grabcut (Rother et al. 2004) to segment the images with three steps: preliminary labels assignment, GMM estimation and energy minimization.

Preliminary Labels Assignment We view the object likelihood values as the prior information for the classification between objects and background. Using $OTSU(P_{co}^{top,n})$, we classify the images into object and background regions. The pixels in object regions are assigned to label 1, and the pixels in background regions are assigned to label 0. For each pixel p,

$$\mathrm{L(p)} = \begin{cases} 1, p \in object \\ 0, p \in background \end{cases} \tag{4.17}$$

GMM Estimation We construct two GMMs, one for object regions and the other for background regions. With the component number $K = 5$, we construct the GMMs based on the preliminary object and background regions, respectively, with the following form:

$$\mathbf{\Omega}(\mathbf{x}_n|\mu_i, \mathbf{\Sigma}_i, \omega_i) = \sum_{i=1}^{K} \omega_i N(\mathbf{x}_n|\mu_i, \mathbf{\Sigma}_i) \tag{4.18}$$

where $\mathbf{x}_n$ is the colour vector of pixel n, K is the number of Gaussian component, $N(\mathbf{x}_n|\mu_i, \mathbf{\Sigma}_i)$ is a Gaussian probability density function, μ_i is the mean vector of data vectors in the same component, $\mathbf{\Sigma}_i$ is the covariance matrix and ω_i is the weight of Gaussian component.

Thus for each pixel p, we can obtain its negative log-likelihood value based

on the two GMMs, respectively,

$$\begin{cases} l_f(p) = -\log(\mathbf{\Omega}_f(p)) \\ l_b(p) = -\log(\mathbf{\Omega}_b(p)) \end{cases} \tag{4.19}$$

Energy Minimization The Gibbs energy for segmentation is defined as:

$$E(L) = U(L) + V(L) \tag{4.20}$$

where $U(L)$ is the data term evaluating the probability of each pixel belonging to object or background,

$$U(L) = \sum_p [l_f(p) \cdot L(p) + l_b(p) \cdot (1 - L(p))] \tag{4.21}$$

and $V(L)$ is the smooth term considering the contrast of neighboured pixels defined as:

$$V(L) = \sum_{(p,q)\in C} [p \neq q]\beta * e^{-DC(p,q)} \tag{4.22}$$

where C is the set of neighborhood pixels, $DC(p, q)$ is the distance between p and q computed using Euclidean distance in the Lab colour space.

By minimizing Equation 4.20, we can obtain new labels of object and background through GMM estimation and energy minimization iteratively.

After obtaining the co-segmentation results of simple images, as shown in Figure 4.1, we extract the samples. The object regions in all layers are extracted as positive samples by assigning the object likelihood as 1. The background regions in all layers are extracted as negative samples by assigning the object likelihood as 0.

4.2.4 Complicated image segmentation

The remaining complicated images are segmented with the guidance of extracted samples which are more reliable than the complicated images themselves.

There are 6 steps to co-segment a remaining complicated image I^n as follows:

step 1) I^n is over-segmented into regions in three hierarchical layers.

step 2) For region $R^{i,l,n}$ in I^n, distance between $R^{i,l,n}$ and all the samples are computed using Equation 4.1 by chi-square distance between their feature histograms. i.e., $R^{i,l,n}$ is compared with all the samples extracted from the N_s simple images.

step 3) A total of N_s most similar samples to $R^{i,l,n}$ are searched using Equation 4.15.

step 4) The common object likelihood value of $R^{i,l,n}$ is calculated using Equation 4.16. If these samples searched in the previous step are object regions (their object likelihood values are assigned 1), then the common object likelihood value of $R^{i,l,n}$ is probably very high; or contrarily, the samples are background regions, then the common object likelihood value of $R^{i,l,n}$ could be very low.

step 5) Equation 4.10 is used to obtain the optimized common object likelihood value of $R^{i,l,n}$.

step 6) Image I^n is segmented through the framework of Grabcut (Boykov & Jolly 2001). We use Equation 4.17 to get preliminary label assignment, and iteratively adopt object/background GMM estimation and Gibbs energy minimization (from Equation 4.18 to Equation 4.22) to obtain the final segmentation results.

Co-segmenting complicated images relying on samples instead of information from the complicated images themselves makes an obvious improvement

on the performance, as shown in Figure 4.5 and Figure 4.1. Compare ranking results in Figure 4.5 with the complicated image co-segmentation results in Figure 4.1: the proposed method can achieve satisfactory segmentation results on the complicated images with low-quality saliency maps, especially for the images in the 3rd row in Figure 4.5.

4.3 Experiments

The experiments are performed on iCoseg dataset (Batra et al. 2010). This dataset contains 38 groups with 643 images in total, and each group has 4 to 41 images. We present the experimental results in both subjective and objective ways, and compare the proposed method with other two state-of-the-art methods (Lee et al. 2015, Faktor & Irani 2013). In order to have a full set of experimental results, a public code of implementing the reference (Lee et al. 2015) has been used, therefore the experimental results in our chapter are not exactly the same as reported in (Lee et al. 2015).

To have an overall performance measure, weighted F-measure, Intersection-over-Union (IoU) and F-measure are used in our experiments. The F-measure score is calculated based on four basic quantities: true-positive (TP), true-negative(TN), false-positive (FP) and false-negative (FN) by

$$\begin{aligned} F_\alpha &= (1+\alpha^2) * \frac{Precision \cdot Recall}{\alpha^2 \cdot Precision + Recall} \\ &= \frac{(1+\alpha^2) \cdot TP}{(1+\alpha^2) \cdot TP + \alpha^2 \cdot FN + FP} \end{aligned} \tag{4.23}$$

In our experiments, we value precision more than recall, and as it is suggested in (Borji et al. 2015), we set $\alpha^2 = 0.3$, the reason for weighting precision more than recall is that recall rate is not as important as precision. For instance, 100% recall can be easily achieved by setting the whole region

to foreground. The weighted F-measure is proposed as a more preferable method to evaluate object segmentation results in (Margolin, Zelnik-Manor & Tal 2014), which replaces the four quantities with weighted quantities and weights errors according to the location and neighborhood.

4.3.1 Parameter setting

In our experiments, we set the number of simple images $N_s = 5$, the maximum region number $M = 100$, the threshold $\theta = 0.7$ for IoU in simple image selection, and the parameter to control the data term and smooth term in energy minimizing $\beta = 0.2$.

4.3.2 Results of the proposed method

To more clearly present the guidance effect of simple images, we show the results from the view of simple/complicated images, and the view of initial/final segmentation in Figure 4.8, which contains all the selected simple images and the most complicated images. We denote RI as ranked images, ISR as initial segmentation results and FSR as final segmentation results. We can find that the object and background regions can be roughly separated already in the initial segmentation results and the final segmentation results show the refinements. The effective initial segmentation is owed to the correct guidance of the simple images and the effectiveness of region matching.

As we mentioned in *Pre-processing*, we use both colour and SIFT features to measure the appearance similarity between regions. These two features can complement each other to some extent. Figure 4.4 has showed some results of region matching, and we can also illustrate the matching performance in Figure 4.8. In "Cheetah", the object regions share higher similarity in



Figure 4.8: Experimental results. In each image group, from top to bottom, RI: Ranked images, ISR: Initial segmentation results, and FSR: Final segmentation results.

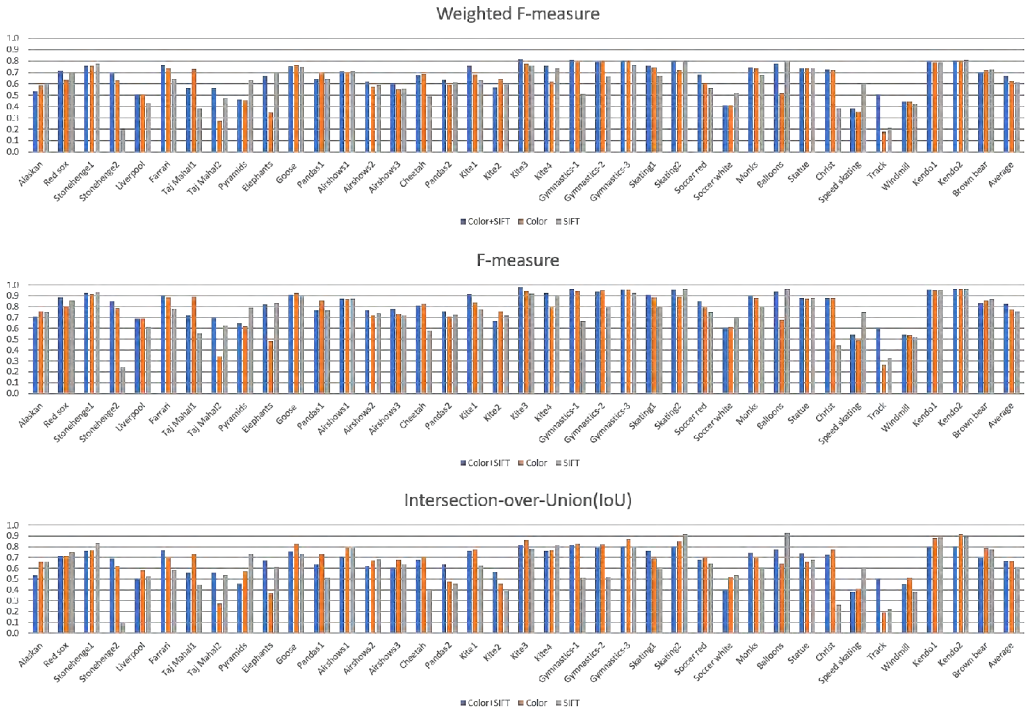

Figure 4.9: Performance on iCoseg Dataset with Different Features.

colour, but in "Stonehenge 2", the objects vary in color due to the change of illumination (e.g. the object in the rightmost image shows different colour from the others). In the case of "Stonehenge 2", the reason why we can segment the object regions is that we also use SIFT feature, and thus the object can be correctly segmented in initial segmentation results. We also compare the results of using both features with the results of using colour or SIFT features alone in Figure 4.9, and apparently, using both features can achieve the better performance.

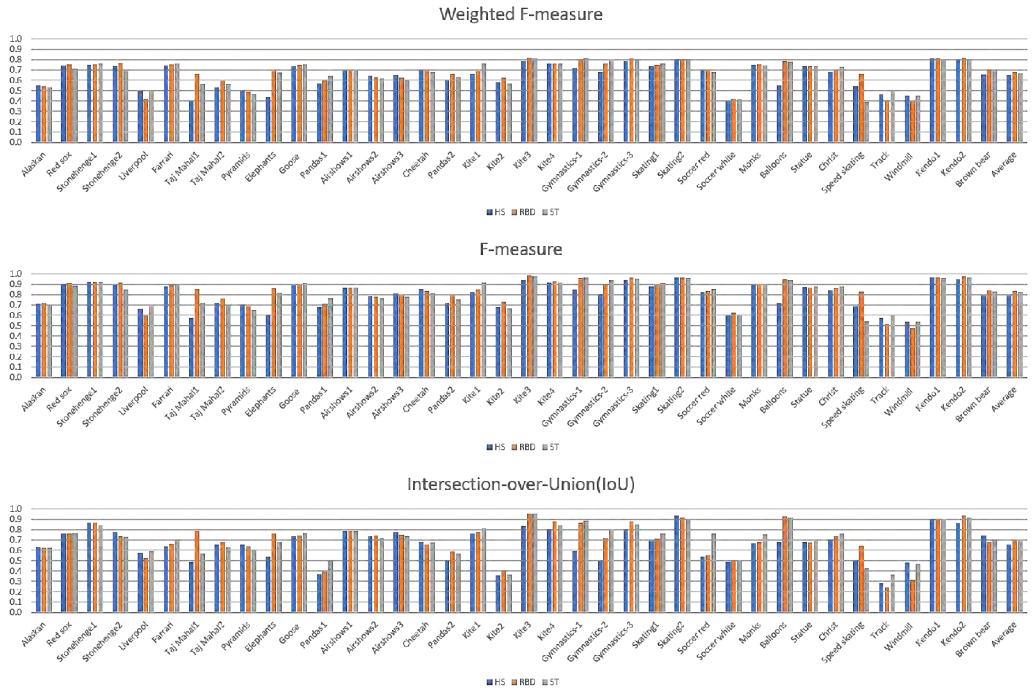

Figure 4.10: Performance on iCoseg Dataset with Different Saliency Models

To objectively evaluate the sensitivity of the proposed method to saliency maps generated using different saliency models, the IoU values, F-measures and weighted F-measures achieved with the three different saliency models including HS (Yan et al. 2013), RBD (Zhu et al. 2014) and ST (Liu et al. 2014) are shown in Figure 4.10. From the average scores in Figure 4.10, RBD

achieves the best both on F-measure and weighted F-measure, ST achieves the best on IoU, while all the three measures with HS are relatively lower compared with RBD and ST. Such a difference with different saliency models is due to the quality of saliency maps. According to the recent benchmark on saliency models (Borji et al. 2015), RBD and ST are within the top six high-performing models, while the performance of HS is lower than RBD and ST. Nonetheless, the average scores of the three measures among the three saliency models do not show substantial differences. Therefore, we can conclude that the proposed method shows the robustness to different saliency models with similar performance, and the better saliency maps generally lead to the better segmentation results.

For simple image groups with simple backgrounds, as Figure 4.11 shows, the proposed method achieves a good performance in the two image groups, and also obtains high scores in the weighted F-measure, IoU and F-measure scores (see "Kendo1" and "Stonehenge1" in Figure 4.14, Figure 4.15 and Figure 4.16).

For the complicated image groups which contain many complicated images or multiple objects as shown in Figure 4.12, the proposed method can obtain the cleaner object regions, e.g. in "Airshow2", the planes can be separated with fewer irrelevant regions ; in group "Elephants", few irrelevant background regions are segmented into the objects.

For image groups in which there are not enough real simple images, and the extracted samples are not correct (i.e., some object regions are extracted as negative samples while some background regions are extracted as positive samples), then the guidance of simple images will not lead to a correct cosegmentation of complicated images. For example, as shown in Figure 4.8 and Figure 4.13, in "Speed skating", background regions (e.g. the advertisement

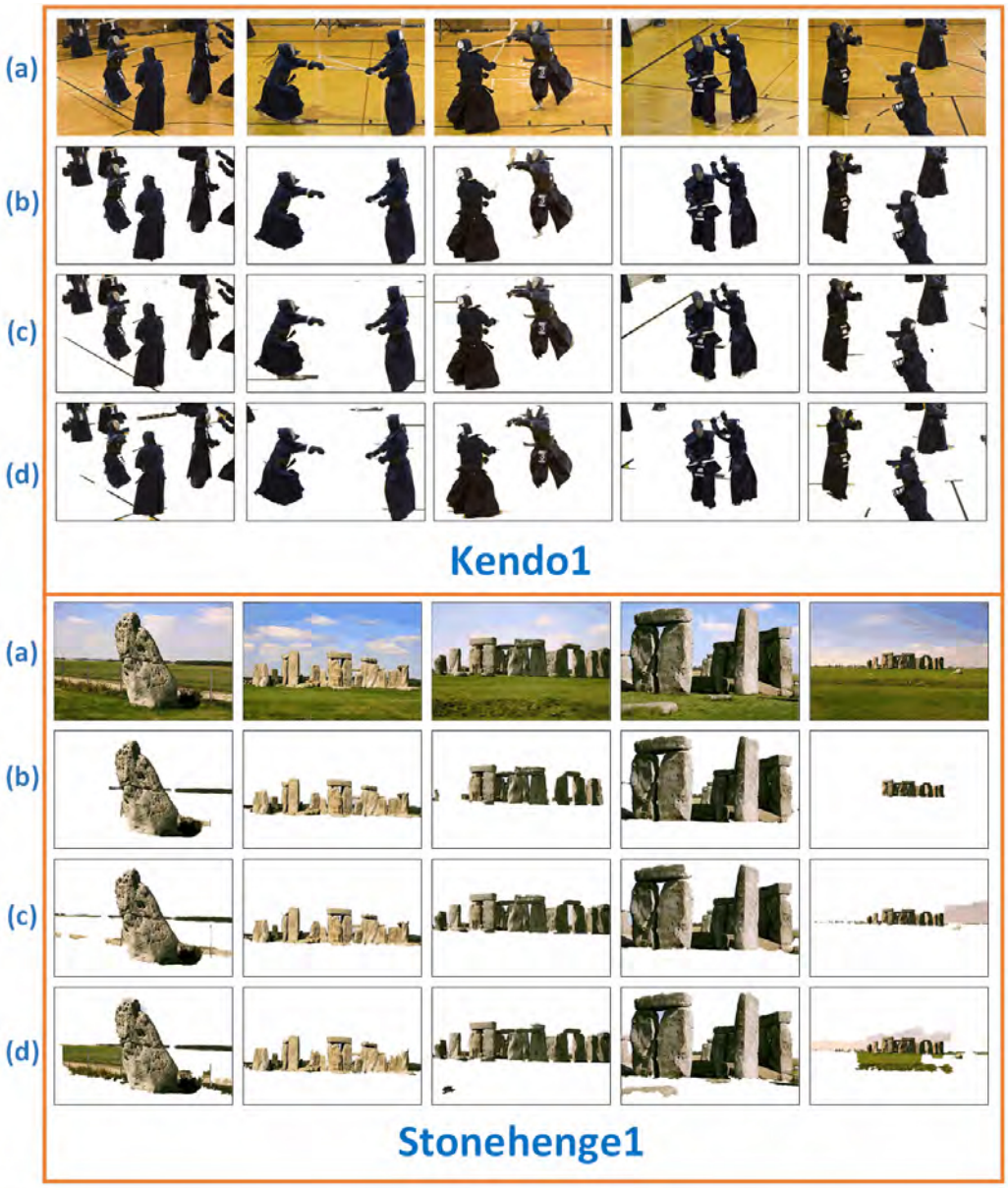

Figure 4.11: Experimental results. (a) the original images, from (b) to (d), the results of the proposed method, (Faktor & Irani 2013) and (Lee et al. 2015), respectively.

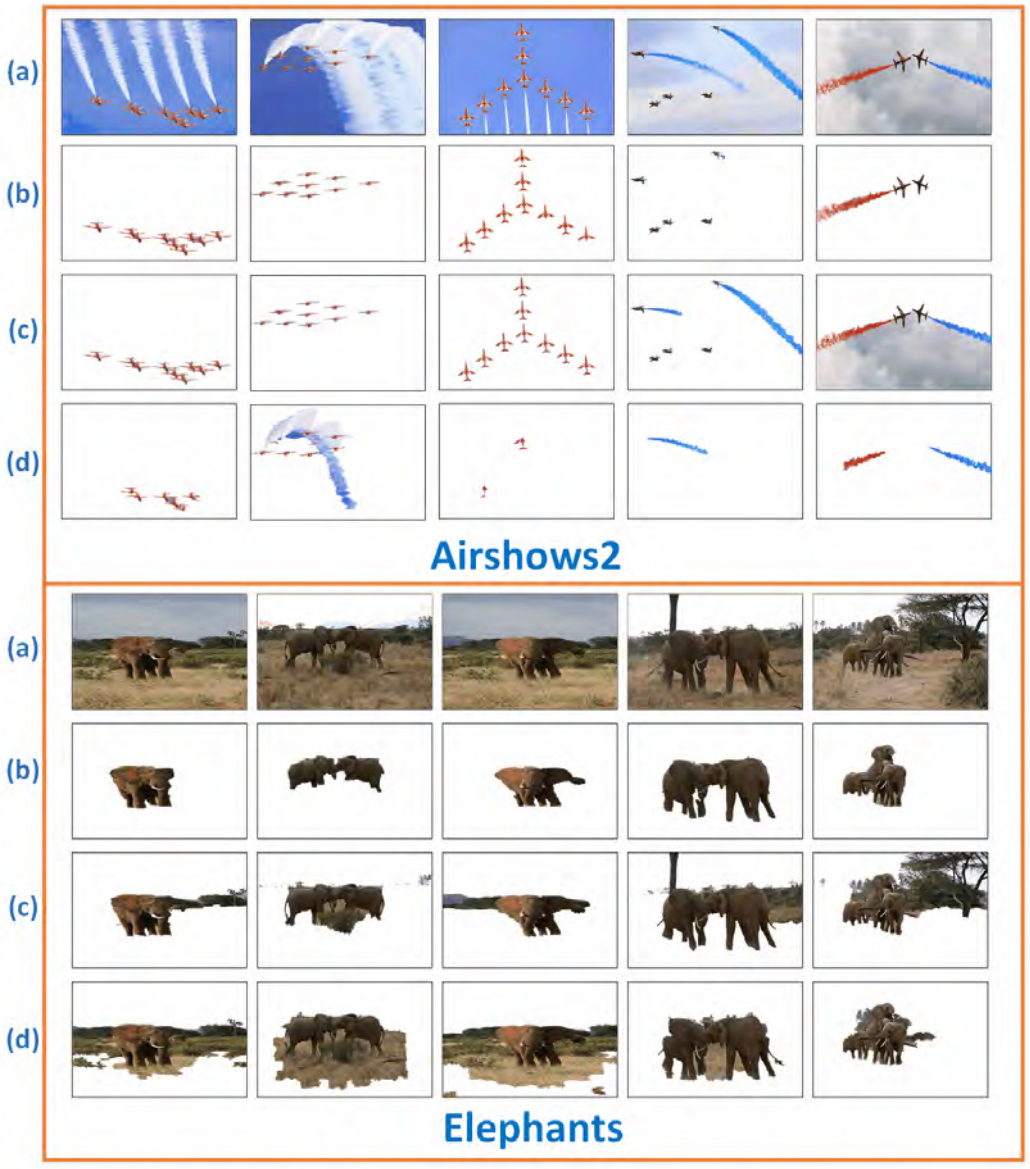

Figure 4.12: Experimental results. (a) the original images, from (b) to (d), the results of the proposed method, (Faktor & Irani 2013) and (Lee et al. 2015), respectively.

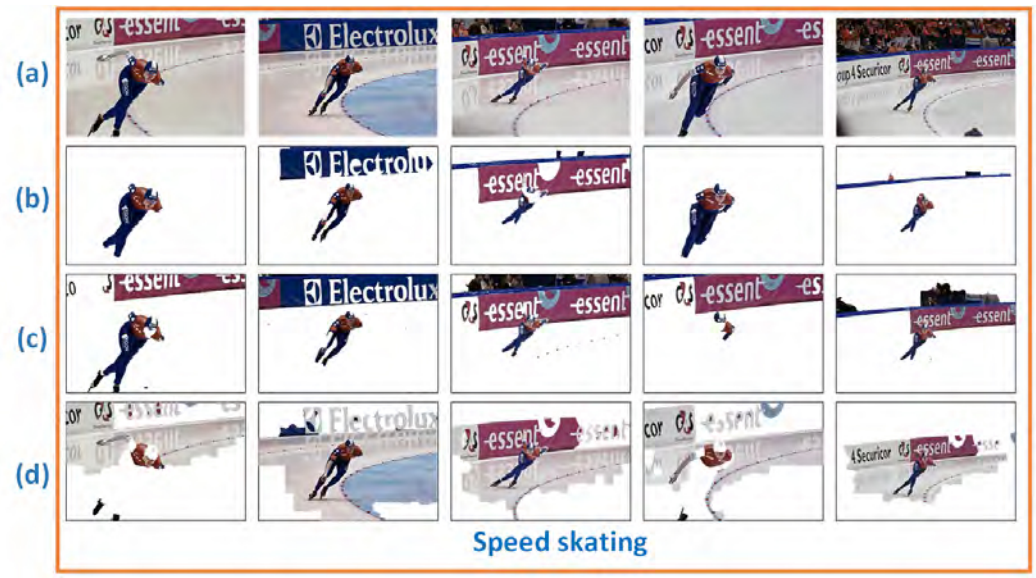

Figure 4.13: Experimental results. (a) the original images, from (b) to (d), the results of the proposed method, (Faktor & Irani 2013) and (Lee et al. 2015), respectively.

sign) in the simple images are extracted as object regions in the initial segmentation results, and then the background regions (e.g. the advertisement sign) in the complicated images are also segmented as object regions.

4.3.3 Comparison with other methods

We compare the proposed method with (Lee et al. 2015, Faktor & Irani 2013), and show some subjective results in Figure 4.11, Figure 4.12 and Figure 4.13, and objective results in Figure 4.14, Figure 4.15 and Figure 4.16. For comparison with (Lee et al. 2015, Faktor & Irani 2013), the saliency model ST (Liu et al. 2014) is used to generate single saliency maps in the proposed method.

For simple image groups as Figure 4.11 shows, the proposed method achieves a good performance on the two image groups with simple back-

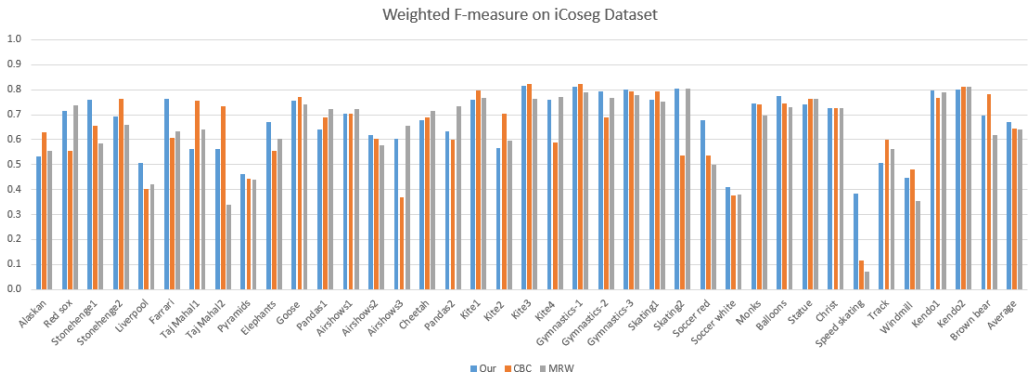

Figure 4.14: Weighted F-measure on iCoseg Dataset

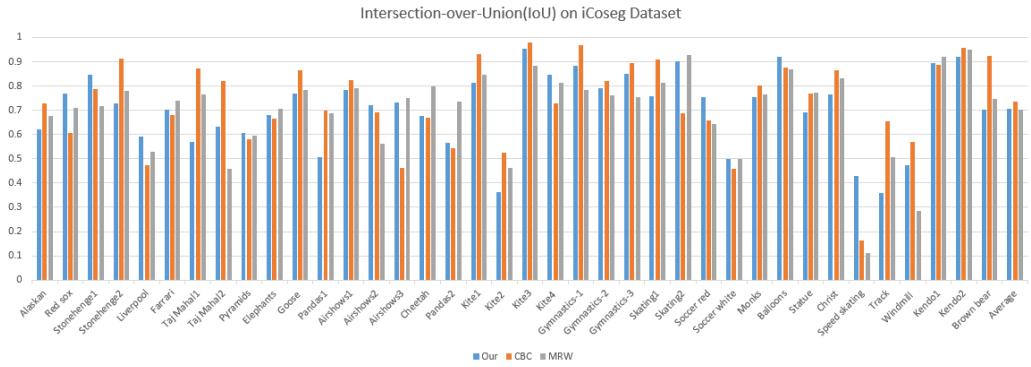

Figure 4.15: Intersection-over-Union(IoU) on iCoseg Dataset

grounds. In "Kendo1", the three methods can achieve satisfactory results. In "Stonehenge1", the results obtained using (Faktor & Irani 2013) and (Lee et al. 2015) contain some irrelevant background regions (e.g. the 1st and 5th columns). The results shown in Figure 4.11 are in consistency with Figure 4.14, Figure 4.15 and Figure 4.16. The three methods can all achieve good segmentation results and high scores on simple image groups.

For complicated image groups, which contain many complicated images, as Figure 4.12 shows, the proposed method can obtain the better results than the other two methods. For example, in "Airshow2", the proposed method can segment the planes with few irrelevant regions; the method (Faktor &

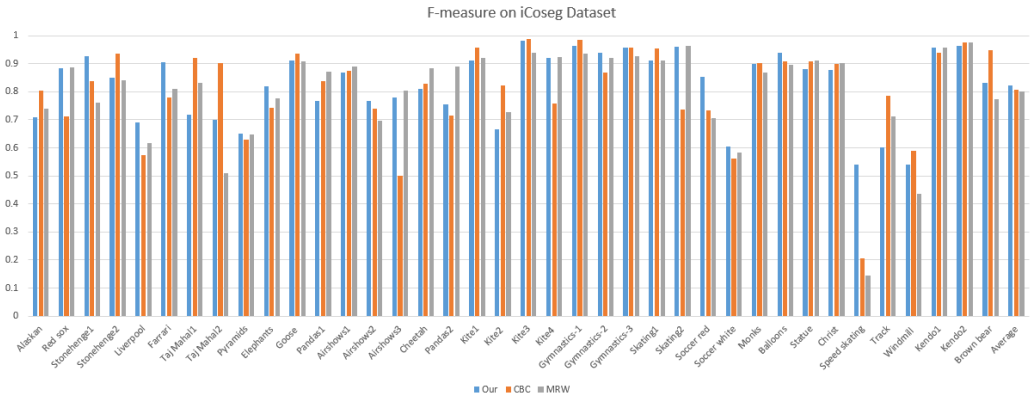

Figure 4.16: F-measure on iCoseg Dataset

Irani 2013) segments the objects with large irrelevant background regions such as those in the 4th and 5th column; and the method (Lee et al. 2015) fails to segment the complete objects. In group "Elephants", our results are clear with few irrelevant background regions while the results of the other two methods contain much more irrelevant regions. The results shown in Figure 4.12 are in consistency with Figure 4.14, Figure 4.15 and Figure 4.16, in which the proposed method achieves the overall higher scores in these groups.

In Figure 4.13, none of the three methods can achieve satisfactory results. Nonetheless, the proposed method achieves the relatively better segmentation quality (e.g. the results in the 1st and 4th column) compared with the other two methods, which totally fail to segment the accurate objects from background in this image group. And for the objective evaluation, the proposed method obtains much higher scores on weighted F-measure, IoU and F-measure.

We can observe from the average scores in Table 4.1 that the proposed method ranks the 1st in terms of F-measure and weighted F-measure, and the 2nd in terms of IoU. In contrast, the method in (Faktor & Irani 2013)

ranks the 2nd in terms of F-measure and weighted F-measure, and the 1st in terms of IoU. Such an observation is due to the relatively higher precision of the proposed method and the relatively higher recall of the method in (Faktor & Irani 2013). Specifically, for the proposed method and the method in (Faktor & Irani 2013), we calculated the average precision as 0.871 and 0.813, respectively, and the average recall as 0.792 and 0.882, respectively. As a comprehensive and overall evaluation using all the three measures, it can be concluded that the proposed method outperforms the other two methods on segmentation performance.

Table 4.1: Overall Performance on iCoseg Dataset

Average Scores	Our	(Faktor & Irani 2013)	(Lee et al. 2015)
Weighted F-measure	0.670	0.645	0.639
F-measure	0.821	0.807	0.800
IoU	0.706	0.734	0.704

4.3.4 Discussion

From the experimental results, we can see that utilizing the information of simple images can truly help improve the performance of image co-segmentation, especially for the complicated images; the colour feature and SIFT feature can complement each other in region matching, and using them together obtains the better performance than using either feature alone. There are also some limitations of the proposed method. If the objects have a very large variance of appearance or there are not enough really simple images (e.g. "Speed skating" in Figure 4.8 and 4.13), the guidance of simple images will be insufficient, and thus the performance of the proposed method may

decrease.

4.4 Summary

In this chapter, we propose a novel image co-segmentation method via simple image guidance. An effective image ranking method is introduced and adopted to rank image complexities based on saliency maps. Then samples are extracted by selecting and segmenting simple images. The complicated images are segmented with the guidance of these samples. The motivation of the proposed method is to utilize the information of simple images instead of the information from all the images. From the experimental results, the proposed method has shown its outperformance by taking advantages of simple images. We believe that this idea can be further explored and be applied not only in image co-segmentation but also other applications.

Chapter 5

Learn Image Object Co-segmentation with Multi-scale Feature Fusion

5.1 Motivations and Significance

There have been quite a few methods for image co-segmentation (Joulin et al. 2010, Rubinstein et al. 2013, Faktor & Irani 2013, Lee et al. 2015, Jerripothula, Cai & Yuan 2018, Li, Liu & Zhang 2018, Tsai et al. 2019, Hsu, Lin & Chuang 2018).

Most of the methods are unsupervised, and the common objects are discovered based on some low-level handcrafted features such as luminance, colors or textures. The similarities among the common objects are computed based on the pixels or regions. In (Joulin et al. 2010), objects are segmented from the background by clustering the image pixels into two clusters that can be maximally distinguished. In (Liu et al. 2013, Li, Liu, Zou, Zhang & Le Meur 2014, Ye, Liu, Li, Zhao & Shen 2015, Faktor & Irani 2013, Lee

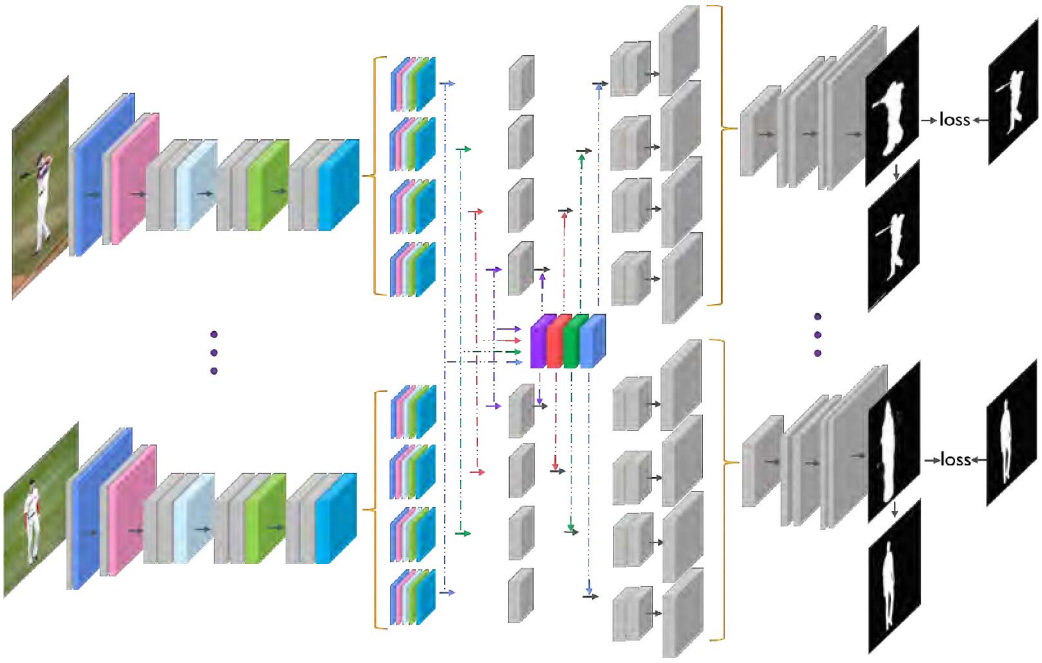

Figure 5.1: An overview of the proposed model. It can be viewed as three parts. The input images are first passed through the reformed VGG networks, which share exactly the same layer structure and parameters. Then, multi-scale [2,4,8,16] features are extracted at each layer (represented in colors and they are stacked according to the scales). They are fused both within each image (in grey) into intra-image features and across images (in colors) into inter-image features, and then intra-image and inter-image features are further fused at each scale, the features at different scales are represented in the corresponding colors. At last, the fused multi-scale features are summed up and upsampled to obtained the coarse co-segmentation results, and final co-segmentation results are obtained via Grabcut refinement.

et al. 2015), regions are compared and computed similarities among all the images of the group while in (Rubinstein et al. 2013), regions are compared only with selected images that share the most similarities between the images, and in (Li, Liu & Zhang 2018), regions are compared only with the selected simple images. In (Jerripothula et al. 2018), guided by quality measurement, saliency maps are fused from different images using the dense correspondence.

Some of the methods are unsupervised but utilizing not only the low-level features but also the high-level semantic features from some pre-trained Convolutional Neural Networks (CNNs). For the recent decade, CNNs have been successfully applied in various research areas and applications especially computer vision, such as image classification, image recognition (Simonyan & Zisserman 2015, He et al. 2016) and image semantic segmentation (Long et al. 2015, Chen, Papandreou, Kokkinos, Murphy & Yuille 2018). In (Zhang, Han, Li & Wang 2015, Tsai et al. 2019), the similarities between regions are computed both by traditional low-level features and high-level CNN features. In (Wei, Zhang, Wu, Shen & Zhou 2019), a Deep Descriptor Transformation (DDT) method is proposed to evaluate the correlations of deep descriptors generated from pre-trained models and then locate the common category-consistent object in a set of unlabeled images. In (Hsu, Lin & Chuang 2018), the common objects are optimized to separated from the backgrounds, to achieve the highest similarities between the objects and lowest similarities between the objects and backgrounds, based on the high-level pre-trained CNNs features.

There are also some methods are fully supervised and trained based on deep CNNs. In (Wei, Zhao, Bourahla, Li, Wu & Zhuang 2019), the networks are trained with shared branches to extract intra-image features for each input image as well as a combined branch to extract inter-image features

(co-features), and the two features are combined to final segmentation predictions. In (Li, Jafari & Rother 2018) the input images are passed through a Siamese encoder to extract feature maps, then feature matching is performed using a mutual correlation network to obtain correspondence maps, and then, the common objects masks are obtained by passing concatenation of squeezed feature maps and correspondence maps through a Siamese decoder.

In image co-segmentation area, deep CNNs have not been fully explored. Thus, in this chapter, a novel neural network with multi-scale feature fusion is proposed. In this network, first, multi-scale convolutional features at multi-layers are extracted through a reformed VGG, then, they are fused across image as the common inter-image features of the objects as well as within each image as intra-image features, and then they are further fused to predict the coarse co-segmentation results. At last, coarse co-segmentation results are refined via grab-cut based method, and the final segmentation results are obtained.

When training a neural network with multiple input images simultaneously, the cost of resources will increase rapidly with the inputs, which raises the requirements of hardware and lowers the training speed. When reducing the resource cost of a CNN, including the volume of the network and computation cost, it is usually considered to 1) reduce the size of input images directly, and if there are down-sampling layers existed, that may lead some feature layers to the degeneration, especially the deeper layers (Krizhevsky et al. 2012, Simonyan & Zisserman 2015, He et al. 2016). 2) reduce the parameters, e.g. kernel sizes, then the performance of the network will be affected. To address the problems we mention above, atrous convolution (Chen, Papandreou, Kokkinos, Murphy & Yuille 2018, Chen et al. 2017) is

adopted in this chapter, which keeps the receptive field with less parameters, together with reduced input size and less downsampling layers.

Overall, the main contributions of the proposed model are two-folders:

- A novel deep neural network with multi-scale feature fusion is proposed for image co-segmentation, and the experimental results demonstrate its comparable performance with the state-of-the-art methods.
- The proposed network has successfully gotten simplified and reduced the cost of resources while achieving comparable results.

The rest of the chapter is organised as follows: in Section 5.2, the details of the proposed model are explained. In Section 5.3, the experimental results are shown, analysed and compared with the state-of-the-art method, and in Section 5.4, the proposed model is concluded.

5.2 The Proposed Model

The proposed model overall contains three parts, as shown in Figure 5.1, the first part is a reformed VGG network (the left to the braces), the second is feature extraction and fusion part (in the braces) in which multi-layer and multi-scale features are extracted and fused, and the third is the upsampling part (the right to the braces) that the coarse co-segmentation results and losses are computed and the final co-segmentation results are refined from the coarse co-segmentation results.

For a group of input images $\{I_1, I_2, ...I_k, ..., I_K\}$, K denotes the number of input images, in this chapter, $K = 5$. First they are passed through the reformed VGG networks (the VGG networks share exactly the same parameters). Then, multi-scale features $F_k^{m,n}, m = \{2, 4, 6, 8\}, n = \{1, 2, ..., 5\}$ are

extracted at each layer (5 layers in total), where k denotes the image/branch, m denotes the scale and n denotes the layer. The features $F_k^{m,n}$ are fused both within each image into intra-image features, denoted as ${F_{intra}}_k^m$, and across images into inter-image features, denoted as ${F_{inter}}^m$, and then intra-image features ${F_{intra}}_k^m$ and inter-image features ${F_{inter}}^m$ are further fused at each scale into co-features ${F_{co}}_k^m$. At last, the fused multi-scale co-features are summed up into F_{cok} and upsampled to obtained the coarse co-segmentation results S_{cok}, and final co-segmentation results R_{cok} are refined from the coarse co-segmentation results S_{cok}.

5.2.1 The reformed VGG network

The reformed VGG network is amended from the standard VGG-16 network. We change the parameters in some layers, and the changing details are shown in Table 5.1, the parameters that not mentioned will stay the same with the original.

Table 5.1: The implementation details of layers in 5.2.1.

layers	**params**	**layers**	**params**
Max pooling 3 Max pooling 4	Kernel: 3×3 Pad: 1 Stride:1	Conv5_2	Dilation: 2 Pad: 2 Output: 256
Conv4_1 Conv4_2 Conv4_3 Conv5_1	Output: 256	Conv5_3	Dilation: 5 Pad: 5 Output: 256

The main purpose to change these layers is that when input images are

far smaller size (such as 128×128), the feature blobs in the network should maintain their functions, thus, the pooling layers and the last two convolutional layers are changed as Table 5.1 shows.

The input images are far small, only the first two pooling layers are kept the same ($stride = 2$), the stride of max pooling 3 and 4 are change into 1 and the feature blob sizes are keep the same without downsampling. To keep the receptive field increasing along the layer depth while the size of parameters is not increasing heavily, dilated convolutional layers are adopted in the last two convolution operation layers, i.e. *Conv5_2* and *Conv5_3*. We also adjust the maximum output number of the layers to 256, as shown in Table 5.1, from the layer *Conv4_1* to *Conv5_3*.

5.2.2 Features extraction and fusion

In this part, multi-layer and multi-scale features are extracted and fused.

As shown in Figure 5.1, after images passed through the reformed VGG, the features at each layer $F_k^{m,n}$ are obtained ($F_k^{m,1}$ on $Conv1_2$ in cyan, $F_k^{m,2}$ on $Conv2_2$ in pink, $F_k^{m,3}$ on $Conv3_3$ in powderblue, $F_k^{m,4}$ on $Conv4_3$ in green, $F_k^{m,5}$ on $Conv5_3$ in blue), and they are then further extracted at multi-scale into ${F_{intra}}_k^m$ and ${F_{inter}}^m$ via dilated convolution, and normalized into the same output size, the details are shown in Table 5.2.

At each scale, the features fused into intra-image features within images as well as into inter-image features across images, to make it clear, the fusion in different scales are represented in different colors (purple, red, dark green and cyan, respectively),as shown in Figure 5.1. Then, the intra-image features and the inter-image features are fused together and upsampled at each scale into multi-scale co-features F_{cok}. The details are shown in Table 5.2.

Table 5.2: The implementation details of the layers in Section 5.2.2.

layers	**params**	**layers**	**params**
Scale1	Kernel: 3×3 Dilation: 2 Pad: 2 Output: 64	Fusion-intra Fusion-inter	Kernel: 1×1 Output: 256
Scale2	Kernel: 3×3 Dilation: 4 Pad: 4 Output: 64	Deconv1_1	Kernel: 1×1 Output: 256
Scale3	Kernel: 3×3 Dilation: 8 Pad: 8 Output: 64	Deconv1_2	Kernel: 3×3 Pad: 1 Output: 256
Scale4	Kernel: 3×3 Dilation: 16 Pad: 16 Output: 64	Upsampling1	Kernel: 2×2 Output: 256

5.2.3 Upsampling

In this part, the multi-scale co-features F_{cok} are summed up and further upsampled into coarse co-segmentation results S_{cok}. The structures of each branch are the same and the parameters are shared. The details are shown in Table 5.3. For the loss function, we use the cross-entropy, the same as in (Chen, Papandreou, Kokkinos, Murphy & Yuille 2018), as loss in each branch to train our network, the difference is that our model simultaneously

have K losses for K inputs correspondingly.

The coarse co-segmentation results S_{cok} may fail to obtain accurate object boundaries, due to the small input size of the training images. To refine the blurry coarse co-segmentation results, we use grab-cut method (Rother et al. 2004) to segment the coarse co-segmentation results into final co-segmentation results R_{cok}, which is one regular method to binarize the network output. Here we use four input images together in grab-cut for each time, and that helps to suppress some common background noise.

Table 5.3: The implementation details of the layers in 5.2.3

layers	**params**	**layers**	**params**
Deconv2_1	Kernel: 3×3 Pad: 1 Output: 128	Deconv3_2	Kernel: 3×3 Pad: 1 Output: 32
Upsampling2	Kernel: 2×2 Output: 128	Prediction	Kernel: 3×3 Pad: 1 Output: 2
Deconv3_1	Kernel: 3×3 Pad: 1 Output: 64		

From the implementation details we listed above in tables 5.1, 5.2, and 5.3, the proposed network is quite simplified and its volume (about $224.6M$ in total) is obviously much smaller than standard setting networks. However, the proposed model can still achieve comparable performance with the simplicity.

5.3 Experiments

5.3.1 Experimental settings

Datasets

To train our network, we use the Cosal2015 dataset (Zhang, Han, Li & Wang 2015) (50 image classes and 2015 images in total), the PASCAL-VOC dataset (Everingham, Van Gool, Williams, Winn & Zisserman 2010) (20 image classes and 1037 images in total) and the Coseg-Rep dataset (Dai, Nian Wu, Zhou & Zhu 2013) (23 image classes and 572 images in total) as the training dataset. To obtain the training image samples that each contains K images, we randomly rank the images belong to a class and then select every K images in order as one training sample, in this chapter, we set $K = 5$ and we finally have 14496 training samples.

And to verify the effectiveness of the proposed image co-segmentation model, we choose the public image co-segmentation dataset (Batra et al. 2010) (38 image classes, and 643 images in total, each class has 4 to 41 images.) as the test dataset.

Implementation details

We implement our model based on the Caffe toolbox (Jia, Shelhamer, Donahue, Karayev, Long, Girshick, Guadarrama & Darrell 2014). The proposed network is initialized randomly. In the training phase, we use the standard stochastic gradient descent (SGD) method with batch size 8, momentum 0.9 and weight decay 0.005. The learning rate is set to be 10^{-6} and keeps fixed. The input size is set to 128×128. The proposed model needs about 80k training iterations for convergence.

Evaluation metrics

To have an overall performance measurement, Intersection-over-Union (IoU) is used in our experiments. IoU is defined as:

$$IoU = \frac{|R_{co} \bigcap GT|}{|R_{co} \bigcup GT|} \tag{5.1}$$

The higher IoU is, the closer the co-segmentation result is to the ground truth, i.e. the better the co-segmentation result is.

5.3.2 Experimental results

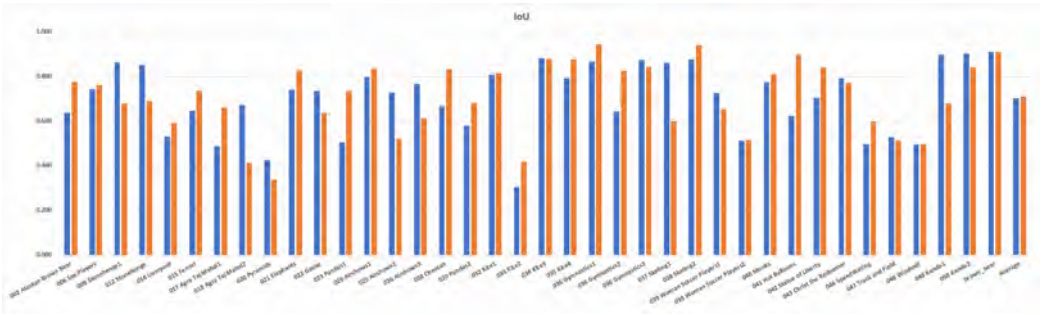

Figure 5.2: Experimental results, objective evaluations.

We present the experimental results in both objective and subjective ways, and compare the proposed model with an recent state-of-the-art model (Li, Liu & Zhang 2018).

The objective evaluation results of IoU performance are shown in Figure 5.2. The proposed model obtains an average score 0.711 on IoU, better than 0.702 on IoU by (Li, Liu & Zhang 2018). Experimental results on this public dataset demonstrate that the proposed model achieves a comparable performance to the state-of-the-art co-segmentation methods.

Some subjective evaluation results are also shown in Figure 5.3 and Figure 5.4. In Figure 5.3, in the group of "skaters", the results of ours contain less

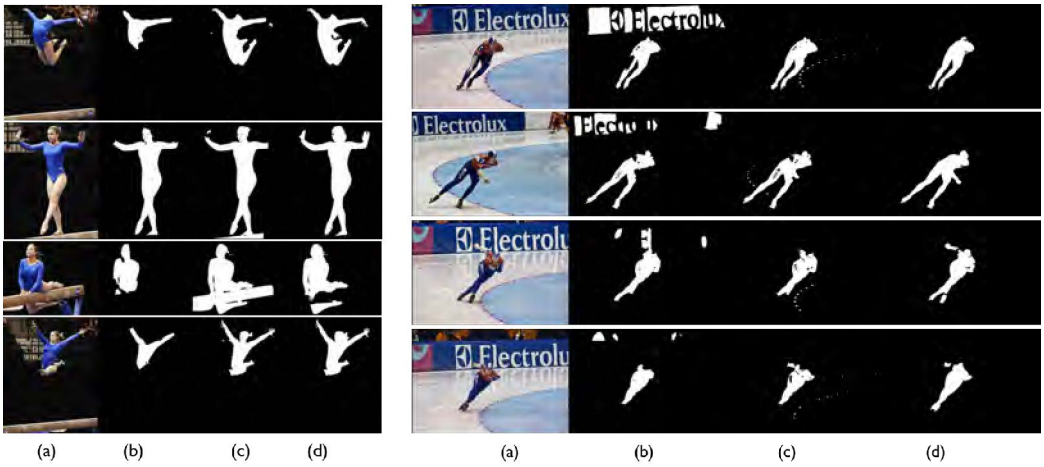

Figure 5.3: Experimental results. (a) the original images, from (b) to (d), the results of (Li, Liu & Zhang 2018), the proposed model, and ground truth, respectively.

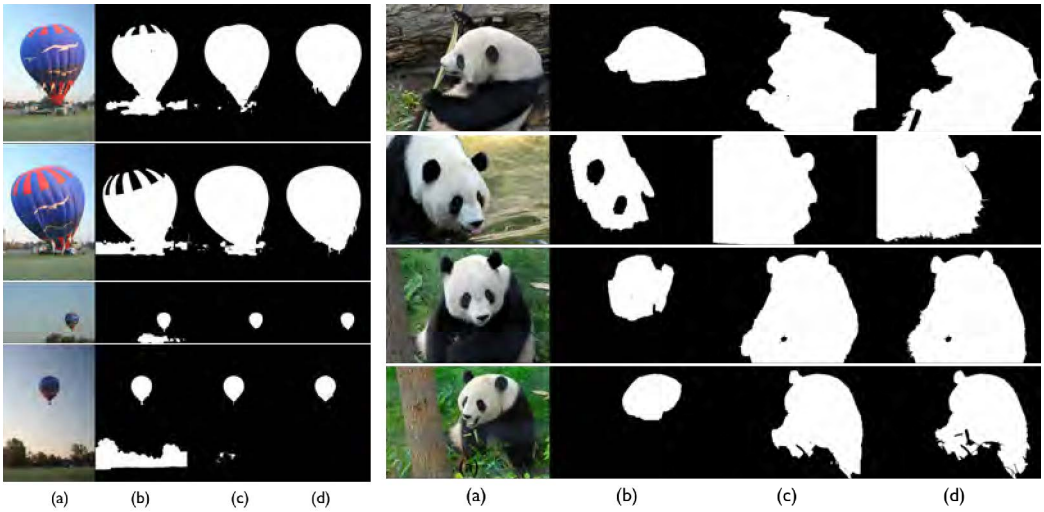

Figure 5.4: Experimental results. (a) the original images, from (b) to (d), the results of (Li, Liu & Zhang 2018), the proposed model, and ground truth, respectively.

background regions compare with (Li, Liu & Zhang 2018), while in "gymnastics" of Figure 5.3 and in "balloons" and "panda" as shown in Figure 5.4, our model segments the common object more complete. We attribute these improvements to the successful utilization of deep network in which concept of whole object is explored better.

5.4 Summary

In this chapter, a novel deep neural network with multi-scale feature fusion is proposed for image co-segmentation, and it has successfully reduced the cost of resource by simplifying the structure (i.e. decreasing the volume of the networks), reducing the input size (decreasing the size of calculation data) while the experimental results demonstrate its comparable performance with the state-of-the-art methods.

Chapter 6

Conclusions and Future Work

6.1 Conclusions

This thesis presents several techniques to address the problem in image co-saliency detection and image co-segmentation.

Chapter 3 proposes an image co-saliency detection model based on region-level fusion and pixel-level refinement. The commonalities between the common objects are defined by the appearance similarities on the regions from all the images. First, the segmentation driven low-rank matrix recovery model is used for intra saliency detection in each individual image of the image set, to highlight the regions whose features are sparse in each image. Then, a region-level fusion method, which exploits inter-region dissimilarities on color histograms and global consistency of regions over the image set, adjusts the intra saliency maps to obtain the region-level co-saliency maps, which can highlight co-salient object regions and suppress irrelevant regions. Finally, a pixel-level refinement method, which integrates color-spatial similarity between pixel and region with image border connectivity based object prior, generates the pixel-level co-saliency maps with better quality. Exten-

sive experiments on two benchmark datasets demonstrate that the proposed co-saliency model consistently outperforms the state-of-the-art co-saliency models in both subjective and objective evaluation.

Chapter 4 proposes an unsupervised images co-segmentation method via guidance of simple images. The commonalities are still defined by handcrafted features on regions. It takes advantages of the reliability of simple images. The images are effectively first ranked by the complexities based on their saliency maps. Then, the simple images, in which objects are common and easy to be segmented, are selected and processed to obtain their segmentation results, these segmentation results are taken as the samples of the targeted objects. Finally, the remaining complicated images are segmented with the guidance of the samples. The experiments on the iCoseg dataset demonstrate the outperformance and robustness of the proposed method.

Chapter 5 proposes a learned image co-segmentation model based on convolutional neural network with multi-scale feature fusion. The commonalities between objects are not defined by handcraft features but learned from the training data. The input images are first passed through the reformed VGG networks. Then, multi-scale [2,4,8,16] features are extracted at each layer. They are fused both within each image into intra-image features and across images into inter-image features, and then intra-image and inter-image features are further fused at each scale, the features at different scales are represented in the corresponding colors. At last, the fused multi-scale features are summed up and upsampled to obtained the coarse co-segmentation results, and final co-segmentation results are obtained by refining the coarse co-segmentation results. Experimental results on the public dataset demonstrate that the proposed model achieves a comparable performance to the state-of-the-art methods while the network has successfully gotten simplified

and the resources cost is reduced.

Chapter 3, Chapter 4 and Chapter 5 of this thesis are supported by the published conference and journal papers listed in List of Publications.

6.2 Future Work

The research on image co-saliency detection and co-segmentation has not been fully explored, especially for learning based co-saliency detection and co-segmentation. The commonalities of common objects have not been unified from the low-level features (colors, textures, etc.) and the high-level semantic features. Based on our work in Chapter 5, the powerful ability of convolutional neural networks on extracting common features can be further utilized, e.g., GAN (Goodfellow, Pouget-Abadie, Mirza, Xu, Warde-Farley, Ozair, Courville & Bengio 2014) has not been used on image co-saliency and co-segmentation research. We believe that the commonalities of common objects should unify the low-level and high-level features together.

Image co-saliency detection and co-segmentation can also be extended to video co-saliency detection and co-segmentation (Chen, Chen & Chang 2012, Wang, Shen, Li & Porikli 2015), which aims to detect and segment the common objects in a video sets. The commonalities of the common objects in videos then contain not only the low-level and high-level spatial features but also the temporal features.

Our future work will still focus on extracting the commonalities of common objects in an image/video set.

Bibliography

Achanta, R., Estrada, F., Wils, P. & Süsstrunk, S. (2008), Salient region detection and segmentation, *in* 'International Conference on Computer Vision Systems', Springer, pp. 66–75.

Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P. & Süsstrunk, S. (2012), 'Slic superpixels compared to state-of-the-art superpixel methods', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **34**(11), 2274–2282.

Amirul Islam, M., Rochan, M., Bruce, N. D. & Wang, Y. (2017), Gated feedback refinement network for dense image labeling, *in* 'Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition', pp. 3751–3759.

Arbelaez, P., Maire, M., Fowlkes, C. & Malik, J. (2011), 'Contour detection and hierarchical image segmentation', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **33**(5), 898–916.

Badrinarayanan, V., Kendall, A. & Cipolla, R. (2017), 'Segnet: A deep convolutional encoder-decoder architecture for image segmentation', *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* **39**(12), 2481–2495.

Bansal, A., Ma, S., Ramanan, D. & Sheikh, Y. (2018), Recycle-gan: Unsupervised video retargeting, *in* 'Proceedings of the European Conference on Computer Vision', pp. 119–135.

Batra, D., Kowdle, A., Parikh, D., Luo, J. & Chen, T. (2010), icoseg: Interactive co-segmentation with intelligent scribble guidance, *in* 'Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition', IEEE, pp. 3169–3176.

Borji, A. (2012), Boosting bottom-up and top-down visual features for saliency estimation, *in* 'Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition', IEEE, pp. 438–445.

Borji, A., Cheng, M.-M., Jiang, H. & Li, J. (2015), 'Salient object detection: A benchmark', *IEEE Transactions on Image Processing* **24**(12), 5706–5722.

Borji, A. & Itti, L. (2012), Exploiting local and global patch rarities for saliency detection, *in* 'Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition', IEEE, pp. 478–485.

Boykov, Y. & Funka-Lea, G. (2006), 'Graph cuts and efficient nd image segmentation', *International Journal of Computer Vision* **70**(2), 109–131.

Boykov, Y. Y. & Jolly, M.-P. (2001), Interactive graph cuts for optimal boundary & region segmentation of objects in nd images, *in* 'Proceedings of the IEEE International Conference on Computer Vision', Vol. 1, IEEE, pp. 105–112.

Bruce, N. & Tsotsos, J. (2006), Saliency based on information maximization, *in* 'Advances in Neural Information Processing Systems', pp. 155–162.

Cao, X., Cheng, Y., Tao, Z. & Fu, H. (2014), Co-saliency detection via base reconstruction, *in* 'Proceedings of the ACM international conference on Multimedia', ACM, pp. 997–1000.

Cao, X., Tao, Z., Zhang, B., Fu, H. & Feng, W. (2014), 'Self-adaptively weighted co-saliency detection via rank constraint', *IEEE Transactions on Image Processing* **23**(9), 4175–4186.

Cao, X., Tao, Z., Zhang, B., Fu, H. & Li, X. (2013), Saliency map fusion based on rank-one constraint, *in* 'Proceedings of the IEEE International Conference on Multimedia and Expo', IEEE, pp. 1–6.

Chang, H.-S. & Wang, Y.-C. F. (2015), 'Optimizing the decomposition for multiple foreground cosegmentation', *Computer Vision and Image Understanding* **141**, 18–27.

Chen, D.-J., Chen, H.-T. & Chang, L.-W. (2012), Video object cosegmentation, *in* 'Proceedings of the ACM International Conference on Multimedia', ACM, pp. 805–808.

Chen, H.-T. (2010), Preattentive co-saliency detection, *in* 'Proceedings of the IEEE International Conference on Image Processing', IEEE, pp. 1117–1120.

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K. & Yuille, A. L. (2018), 'Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **40**(4), 834–848.

Chen, L.-C., Papandreou, G., Schroff, F. & Adam, H. (2017), 'Rethinking atrous convolution for semantic image segmentation', *arXiv preprint arXiv:1706.05587* .

Chen, L.-C., Yang, Y., Wang, J., Xu, W. & Yuille, A. L. (2016), Attention to scale: Scale-aware semantic image segmentation, *in* 'Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition', pp. 3640–3649.

Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F. & Adam, H. (2018), Encoder-decoder with atrous separable convolution for semantic image segmentation, *in* 'Proceedings of the European Conference on Computer Vision', pp. 801–818.

Chen, M., Velasco-Forero, S., Tsang, I. & Cham, T.-J. (2015), Objects co-segmentation: Propagated from simpler images, *in* 'Proceedings of the Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing', IEEE, pp. 1682–1686.

Cheng, M.-M., Mitra, N. J., Huang, X. & Hu, S.-M. (2014), 'Salientshape: Group saliency in image collections', *The Visual Computer* **30**(4), 443–453.

Cheng, M.-M., Mitra, N. J., Huang, X., Torr, P. H. & Hu, S.-M. (2014), 'Global contrast based salient region detection', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **37**(3), 569–582.

Cheng, M.-M., Warrell, J., Lin, W.-Y., Zheng, S., Vineet, V. & Crook, N. (2013), Efficient salient region detection with soft image abstraction, *in* 'Proceedings of the IEEE International Conference on Computer Vision', pp. 1529–1536.

Cheriet, M., Said, J. N. & Suen, C. Y. (1998), 'A recursive thresholding technique for image segmentation', *IEEE Transactions on Image Processing* **7**(6), 918–921.

Cho, D., Park, J., Oh, T.-H., Tai, Y.-W. & So Kweon, I. (2017), Weakly- and self-supervised learning for content-aware deep image retargeting, *in* 'Proceedings of the IEEE International Conference on Computer Vision', pp. 4558–4567.

Cho, M., Kwak, S., Schmid, C. & Ponce, J. (2015), 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', pp. 1201–1210.

Cho, M., Shin, Y. M. & Lee, K. M. (2008), Co-recognition of image pairs by data-driven monte carlo image exploration, *in* 'Proceedings of the European Conference on Computer Vision', Springer, pp. 144–157.

Choi, H. & Baraniuk, R. G. (2001), 'Multiscale image segmentation using wavelet-domain hidden markov models', *IEEE Transactions on Image Processing* **10**(9), 1309–1321.

Chollet, F. (2017), Xception: Deep learning with depthwise separable convolutions, *in* 'Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition', pp. 1251–1258.

Chuang, K.-S., Tzeng, H.-L., Chen, S., Wu, J. & Chen, T.-J. (2006), 'Fuzzy c-means clustering with spatial information for image segmentation', *Computerized Medical Imaging and Graphics* **30**(1), 9–15.

Comaniciu, D. & Meer, P. (2002), 'Mean shift: A robust approach toward feature space analysis', *IEEE Transactions on Pattern Analysis and Machine Intelligence* (5), 603–619.

Cong, R., Lei, J., Fu, H., Huang, Q., Cao, X. & Hou, C. (2017), 'Co-saliency detection for rgbd images based on multi-constraint feature matching

and cross label propagation', *IEEE Transactions on Image Processing* **27**(2), 568–579.

Dai, J., He, K. & Sun, J. (2016), Instance-aware semantic segmentation via multi-task network cascades, *in* 'Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition', pp. 3150–3158.

Dai, J., Li, Y., He, K. & Sun, J. (2016), R-fcn: Object detection via region-based fully convolutional networks, *in* 'Advances in Neural Information Processing Systems', pp. 379–387.

Dai, J., Nian Wu, Y., Zhou, J. & Zhu, S.-C. (2013), Cosegmentation and cosketch by unsupervised learning, *in* 'Proceedings of the IEEE International Conference on Computer Vision', pp. 1305–1312.

Everingham, M., Van Gool, L., Williams, C. K., Winn, J. & Zisserman, A. (2010), 'The pascal visual object classes (voc) challenge', *International Journal of Computer Vision* **88**(2), 303–338.

Faktor, A. & Irani, M. (2013), Co-segmentation by composition, *in* 'Proceedings of the IEEE International Conference on Computer vision', IEEE, pp. 1297–1304.

Felzenszwalb, P. F. & Huttenlocher, D. P. (2004), 'Efficient graph-based image segmentation', *International Journal of Computer Vision* **59**(2), 167–181.

Fourure, D., Emonet, R., Fromont, E., Muselet, D., Tremeau, A. & Wolf, C. (2017), 'Residual conv-deconv grid network for semantic segmentation', *arXiv preprint arXiv:1707.07958* .

Freixenet, J., Muñoz, X., Raba, D., Martí, J. & Cufí, X. (2002), Yet another survey on image segmentation: Region and boundary information integration, *in* 'Proceedings of the European Conference on Computer Vision', Springer, pp. 408–422.

Fu, H., Cao, X. & Tu, Z. (2013), 'Cluster-based co-saliency detection', *IEEE Transactions on Image Processing* **22**(10), 3766–3778.

Fu, J., Liu, J., Wang, Y., Zhou, J., Wang, C. & Lu, H. (2019), 'Stacked deconvolutional network for semantic segmentation', *IEEE Transactions on Image Processing* .

Fulkerson, B., Vedaldi, A. & Soatto, S. (2009), Class segmentation and object localization with superpixel neighborhoods, *in* 'Proceedings of the IEEE International Conference on Computer Vision', IEEE, pp. 670–677.

Gao, Y., Wang, M., Tao, D., Ji, R. & Dai, Q. (2012), '3-d object retrieval and recognition with hypergraph analysis', *IEEE Transactions on Image Processing* **21**(9), 4290–4303.

Ge, C., Fu, K., Liu, F., Bai, L. & Yang, J. (2016), 'Co-saliency detection via inter and intra saliency propagation', *Signal Processing: Image Communication* **44**, 69–83.

Goferman, S., Zelnik-Manor, L. & Tal, A. (2011), 'Context-aware saliency detection', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **34**(10), 1915–1926.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. & Bengio, Y. (2014), Generative adversarial nets, *in* 'Advances in Neural Information Processing Systems', pp. 2672–2680.

Guo, C. & Zhang, L. (2009), 'A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression', *IEEE Transactions on Image Processing* **19**(1), 185–198.

Hammouche, K., Diaf, M. & Siarry, P. (2008), 'A multilevel automatic thresholding method based on a genetic algorithm for a fast image segmentation', *Computer Vision and Image Understanding* **109**(2), 163–175.

Han, J., Cheng, G., Li, Z. & Zhang, D. (2017), 'A unified metric learning-based framework for co-saliency detection', *IEEE Transactions on Circuits and Systems for Video Technology* **28**(10), 2473–2483.

Harel, J., Koch, C. & Perona, P. (2007), Graph-based visual saliency, *in* 'Advances in Neural Information Processing Systems', pp. 545–552.

He, K., Zhang, X., Ren, S. & Sun, J. (2015), 'Spatial pyramid pooling in deep convolutional networks for visual recognition', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **37**(9), 1904–1916.

He, K., Zhang, X., Ren, S. & Sun, J. (2016), Deep residual learning for image recognition, *in* 'Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition', pp. 770–778.

Hou, Q., Cheng, M.-M., Hu, X., Borji, A., Tu, Z. & Torr, P. H. (2017), Deeply supervised salient object detection with short connections, *in* 'Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition', pp. 3203–3212.

Hou, X. & Zhang, L. (2007), Saliency detection: A spectral residual approach, *in* 'Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition', Ieee, pp. 1–8.

Hsu, K.-J., Lin, Y.-Y. & Chuang, Y.-Y. (2018), Co-attention cnns for unsupervised object co-segmentation., *in* 'International Joint Conferences on Artificial Intelligence', pp. 748–756.

Hsu, K.-J., Tsai, C.-C., Lin, Y.-Y., Qian, X. & Chuang, Y.-Y. (2018), Unsupervised cnn-based co-saliency detection with graphical optimization, *in* 'Proceedings of the European Conference on Computer Vision', pp. 485–501.

Hu, P., Shuai, B., Liu, J. & Wang, G. (2017), Deep level sets for salient object detection, *in* 'Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition', pp. 2300–2309.

Hu, Y., Rajan, D. & Chia, L.-T. (2005), Robust subspace analysis for detecting visual attention regions in images, *in* 'Proceedings of the ACM International Conference on Multimedia', ACM, pp. 716–724.

Huang, G., Liu, Z., Van Der Maaten, L. & Weinberger, K. Q. (2017), Densely connected convolutional networks, *in* 'Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition', pp. 4700–4708.

Huang, R., Feng, W. & Sun, J. (2015), Saliency and co-saliency detection by low-rank multiscale fusion, *in* 'Proceedings of the IEEE International Conference on Multimedia and Expo', IEEE, pp. 1–6.

Itti, L., Koch, C. & Niebur, E. (1998), 'A model of saliency-based visual attention for rapid scene analysis', *IEEE Transactions on Pattern Analysis and Machine Intelligence* (11), 1254–1259.

Jacobs, D. E., Goldman, D. B. & Shechtman, E. (2010), Cosaliency: Where people look when comparing images, *in* 'Proceedings of the ACM symposium on user interface software and technology', ACM, pp. 219–228.

Jerripothula, K. R., Cai, J. & Yuan, J. (2018), 'Quality-guided fusion-based co-saliency estimation for image co-segmentation and colocalization', *IEEE Transactions on Multimedia* **20**(9), 2466–2477.

Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S. & Darrell, T. (2014), Caffe: Convolutional architecture for fast feature embedding, *in* 'Proceedings of the ACM International Conference on Multimedia', ACM, pp. 675–678.

Jiang, H., Wang, J., Yuan, Z., Liu, T., Zheng, N. & Li, S. (2011), Automatic salient object segmentation based on context and shape prior., *in* 'Proceedings of the British Machine Vision Conference', Vol. 6, p. 9.

Jiang, Z. & Davis, L. S. (2013), Submodular salient region detection, *in* 'Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition', pp. 2043–2050.

Jing, H., He, X., Han, Q. & Niu, X. (2015), 'Co-saliency detection linearly combining single-view saliency and foreground correspondence', *IEICE Transactions on Information and Systems* **98**(4), 985–988.

Joulin, A., Bach, F. & Ponce, J. (2010), Discriminative clustering for image co-segmentation, *in* '2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition', IEEE, pp. 1943–1950.

Joulin, A., Bach, F. & Ponce, J. (2012), Multi-class cosegmentation, *in* 'Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition', IEEE, pp. 542–549.

Judd, T., Ehinger, K., Durand, F. & Torralba, A. (2009), Learning to predict where humans look, *in* 'Proceedings of the IEEE International Conference on Computer Vision', IEEE, pp. 2106–2113.

Kim, G. & Xing, E. P. (2012), On multiple foreground cosegmentation, *in* 'Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition', IEEE, pp. 837–844.

Klein, D. A. & Frintrop, S. (2011), Center-surround divergence of feature statistics for salient object detection, *in* 'Proceedings of the IEEE International Conference on Computer Vision', IEEE, pp. 2214–2219.

Krizhevsky, A., Sutskever, I. & Hinton, G. E. (2012), Imagenet classification with deep convolutional neural networks, *in* 'Advances in Neural Information Processing Systems', pp. 1097–1105.

Kruthiventi, S. S., Gudisa, V., Dholakiya, J. H. & Venkatesh Babu, R. (2016), Saliency unified: A deep architecture for simultaneous eye fixation prediction and salient object segmentation, *in* 'Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition', pp. 5781–5790.

Kuen, J., Wang, Z. & Wang, G. (2016), Recurrent attentional networks for saliency detection, *in* 'Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition', pp. 3668–3677.

Lee, C., Jang, W.-D., Sim, J.-Y. & Kim, C.-S. (2015), Multiple random walkers and their application to image cosegmentation, *in* 'Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition', pp. 3837–3845.

Levinshtein, A., Stere, A., Kutulakos, K. N., Fleet, D. J., Dickinson, S. J. & Siddiqi, K. (2009), 'Turbopixels: Fast superpixels using geometric flows', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **31**(12), 2290–2297.

Li, G., Xie, Y., Lin, L. & Yu, Y. (2017), Instance-level salient object segmentation, *in* 'Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition', pp. 2386–2395.

Li, G. & Yu, Y. (2015), Visual saliency based on multiscale deep features, *in* 'Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition', pp. 5455–5463.

Li, G. & Yu, Y. (2016), Deep contrast learning for salient object detection, *in* 'Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition', pp. 478–487.

Li, H. & Ngan, K. N. (2011), 'A co-saliency model of image pairs', *IEEE Transactions on Image Processing* **20**(12), 3365–3375.

Li, L., Liu, Z. & Zhang, J. (2018), 'Unsupervised image co-segmentation via guidance of simple images', *Neurocomputing* **275**, 1650–1661.

Li, L., Liu, Z., Zou, W., Zhang, X. & Le Meur, O. (2014), Co-saliency detection based on region-level fusion and pixel-level refinement, *in* 'Proceedings of the IEEE International Conference on Multimedia and Expo', IEEE, pp. 1–6.

Li, M., Zuo, W., Gu, S., Zhao, D. & Zhang, D. (2018), Learning convolutional networks for content-weighted image compression, *in* 'Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition', pp. 3214–3223.

Li, W., Jafari, O. H. & Rother, C. (2018), Deep object co-segmentation, *in* 'Proceedings of the Asian Conference on Computer Vision', Springer, pp. 638–653.

Li, X., Li, Y., Shen, C., Dick, A. & Van Den Hengel, A. (2013), Contextual hypergraph modeling for salient object detection, *in* 'Proceedings of the IEEE International Conference on Computer Vision', pp. 3328–3335.

Li, X., Lu, H., Zhang, L., Ruan, X. & Yang, M.-H. (2013), Saliency detection via dense and sparse reconstruction, *in* 'Proceedings of the IEEE International Conference on Computer Vision', pp. 2976–2983.

Li, Y., Liu, J., Li, Z., Lu, H. & Ma, S. (2016), 'Object co-segmentation via salient and common regions discovery', *Neurocomputing* **172**, 225–234.

Liang, Y., Liu, Y.-J. & Gutierrez, D. (2016), 'Objective quality prediction of image retargeting algorithms', *IEEE Transactions on Visualization and Computer Graphics* **23**(2), 1099–1110.

Lin, B.-C., Chen, D.-J. & Chang, L.-W. (2015), Unsupervised image co-segmentation based on cooperative game, *in* 'Proceedings of the Asian Conference on Computer Vision', Springer, pp. 51–63.

Lin, G., Milan, A., Shen, C. & Reid, I. (2017), Refinenet: Multi-path refinement networks for high-resolution semantic segmentation, *in* 'Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition', pp. 1925–1934.

Lin, S.-S., Yeh, I.-C., Lin, C.-H. & Lee, T.-Y. (2013), 'Patch-based image warping for content-aware retargeting', *IEEE Transactions on Multimedia* **15**(2), 359–368.

Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B. & Belongie, S. (2017), Feature pyramid networks for object detection, *in* 'Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition', pp. 2117–2125.

Liu, F. & Gleicher, M. (2006), Region enhanced scale-invariant saliency detection, *in* 'Proceedings of the IEEE International Conference on Multimedia and Expo', IEEE, pp. 1477–1480.

Liu, F., Lin, G. & Shen, C. (2015), 'Crf learning with cnn features for image segmentation', *Pattern Recognition* **48**(10), 2983–2992.

Liu, G.-H. & Yang, J.-Y. (2013), 'Content-based image retrieval using color difference histogram', *Pattern Recognition* **46**(1), 188–198.

Liu, J., Wang, B., Li, M., Li, Z., Ma, W., Lu, H. & Ma, S. (2007), Dual cross-media relevance model for image annotation, *in* 'Proceedings of the ACM International Conference on Multimedia', ACM, pp. 605–614.

Liu, N. & Han, J. (2016), Dhsnet: Deep hierarchical saliency network for salient object detection, *in* 'Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition', pp. 678–686.

Liu, N. & Han, J. (2018), 'A deep spatial contextual long-term recurrent convolutional network for saliency detection', *IEEE Transactions on Image Processing* **27**(7), 3264–3274.

Liu, N., Han, J. & Yang, M.-H. (2018), Picanet: Learning pixel-wise contextual attention for saliency detection, *in* 'Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition', pp. 3089–3098.

Liu, N., Han, J., Zhang, D., Wen, S. & Liu, T. (2015), Predicting eye fixations using convolutional neural networks, *in* 'Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition', pp. 362–370.

Liu, T., Yuan, Z., Sun, J., Wang, J., Zheng, N., Tang, X. & Shum, H.-Y. (2010), 'Learning to detect a salient object', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **33**(2), 353–367.

Liu, Z., Shi, R., Shen, L., Xue, Y., Ngan, K. N. & Zhang, Z. (2012), 'Unsupervised salient object segmentation based on kernel density estimation and two-phase graph cut', *IEEE Transactions on Multimedia* **14**(4), 1275–1289.

Liu, Z., Zou, W. & Le Meur, O. (2014), 'Saliency tree: A novel saliency detection framework', *IEEE Transactions on Image Processing* **23**(5), 1937–1952.

Liu, Z., Zou, W., Li, L., Shen, L. & Le Meur, O. (2013), 'Co-saliency detection based on hierarchical segmentation', *IEEE Signal Processing Letters* **21**(1), 88–92.

Lloyd, S. (1982), 'Least squares quantization in pcm', *IEEE Transactions on Information Theory* **28**(2), 129–137.

Long, J., Shelhamer, E. & Darrell, T. (2015), Fully convolutional networks for semantic segmentation, *in* 'Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition', pp. 3431–3440.

Lou, J., Xu, F., Xia, Q., Yang, W. & Ren, M. (2017), Hierarchical co-salient object detection via color names, *in* 'Proceedings of the Asian Conference on Pattern Recognition', IEEE, pp. 718–724.

Luo, Z., Mishra, A., Achkar, A., Eichel, J., Li, S. & Jodoin, P.-M. (2017), Non-local deep features for salient object detection, *in* 'Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition', pp. 6609–6617.

Mai, L. & Liu, F. (2014), Comparing salient object detection results without ground truth, *in* 'Proceedings of the European Conference on Computer Vision', Springer, pp. 76–91.

Margolin, R., Tal, A. & Zelnik-Manor, L. (2013), What makes a patch distinct?, *in* 'Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition', pp. 1139–1146.

Margolin, R., Zelnik-Manor, L. & Tal, A. (2014), How to evaluate foreground maps?, *in* 'Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition', pp. 248–255.

Meng, F., Li, H., Ngan, K. N., Zeng, L. & Wu, Q. (2013), 'Feature adaptive co-segmentation by complexity awareness', *IEEE Transactions on Image Processing* **22**(12), 4809–4824.

Newell, A., Yang, K. & Deng, J. (2016), Stacked hourglass networks for human pose estimation, *in* 'Proceedings of the European Conference on Computer Vision', Springer, pp. 483–499.

Noh, H., Hong, S. & Han, B. (2015), Learning deconvolution network for semantic segmentation, *in* 'Proceedings of the IEEE International Conference on Computer Vision', pp. 1520–1528.

Oh, S. J., Benenson, R., Khoreva, A., Akata, Z., Fritz, M. & Schiele, B. (2017), Exploiting saliency for object segmentation from image level labels, *in* 'Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition', IEEE, pp. 5038–5047.

Otsu, N. (1975), 'A threshold selection method from gray-level histograms', *Automatica* **11**(285-296), 23–27.

Pappas, T. N. (1992), 'An adaptive clustering algorithm for image segmentation', *IEEE Transactions on Signal Processing* **40**(4), 901–914.

Peng, H., Li, B., Ji, R., Hu, W., Xiong, W. & Lang, C. (2013), Salient object detection via low-rank and structured sparse matrix decomposition, *in* 'Proceedings of the AAAI Conference on Artificial Intelligence'.

Pohlen, T., Hermans, A., Mathias, M. & Leibe, B. (2017), Full-resolution residual networks for semantic segmentation in street scenes, *in* 'Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition', pp. 4151–4160.

Porter, R. & Canagarajah, N. (1996), 'A robust automatic clustering scheme for image segmentation using wavelets', *IEEE Transactions on Image Processing* **5**(4), 662–665.

Quan, R., Han, J., Zhang, D. & Nie, F. (2016), Object co-segmentation via graph optimized-flexible manifold ranking, *in* 'Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition', pp. 687–695.

Redmon, J., Divvala, S., Girshick, R. & Farhadi, A. (2016), You only look once: Unified, real-time object detection, *in* 'Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition', pp. 779–788.

Ren, S., He, K., Girshick, R. & Sun, J. (2015), Faster r-cnn: Towards real-time object detection with region proposal networks, *in* 'Advances in Neural Information Processing Systems', pp. 91–99.

Ronneberger, O., Fischer, P. & Brox, T. (2015), U-net: Convolutional networks for biomedical image segmentation, *in* 'Proceedings of the International Conference on Medical Image Computing and Computer-assisted Intervention', Springer, pp. 234–241.

Rosin, P. L. (2009), 'A simple method for detecting salient regions', *Pattern Recognition* **42**(11), 2363–2371.

Rother, C., Kolmogorov, V. & Blake, A. (2004), Grabcut: Interactive foreground extraction using iterated graph cuts, *in* 'ACM Transactions on Graphics', Vol. 23, ACM, pp. 309–314.

Rother, C., Minka, T., Blake, A. & Kolmogorov, V. (2006), 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', Vol. 1, IEEE, pp. 993–1000.

Rubinstein, M., Joulin, A., Kopf, J. & Liu, C. (2013), Unsupervised joint object discovery and segmentation in internet images, *in* 'Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition', pp. 1939–1946.

Scharfenberger, C., Wong, A., Fergani, K., Zelek, J. S. & Clausi, D. A. (2013), Statistical textural distinctiveness for salient region detection in natural images, *in* 'Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition', pp. 979–986.

Setlur, V., Lechner, T., Nienhaus, M. & Gooch, B. (2007), 'Retargeting images and video for preserving information saliency', *IEEE Computer Graphics and Applications* **27**(5), 80–88.

Shamir, A. & Sorkine, O. (2009), Visual media retargeting, *in* 'ACM SIGGRAPH ASIA 2009 Courses', ACM, p. 11.

Shen, L., Liu, Z. & Zhang, Z. (2013), 'A novel H.264 rate control algorithm with consideration of visual attention', *Multimedia Tools and Applications* **63**(3), 709–727.

Shen, N.-M., Li, J., Zhou, P.-Y., Huo, Y. & Zhuang, Y. (2016), 'Bsfcos: Block and sparse principal component analysis-based fast co-saliency detection method', *International Journal of Pattern Recognition and Artificial Intelligence* **30**(01), 1655003.

Shen, X. & Wu, Y. (2012), A unified approach to salient object detection via low rank matrix recovery, *in* 'Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition', IEEE, pp. 853–860.

Shi, J. & Malik, J. (2000), 'Normalized cuts and image segmentation', *Departmental Papers (CIS)* p. 107.

Simonyan, K. & Zisserman, A. (2015), Very deep convolutional networks for large-scale image recognition, *in* 'Proceedings of the International Conference on Learning Representations'.

Szegedy, C., Ioffe, S., Vanhoucke, V. & Alemi, A. A. (2017), Inception-v4, inception-resnet and the impact of residual connections on learning, *in* 'Proceedings of the AAAI Conference on Artificial Intelligence'.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. & Wojna, Z. (2016), Rethinking the inception architecture for computer vision, *in* 'Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition', pp. 2818–2826.

Tan, Z., Wan, L., Feng, W. & Pun, C.-M. (2013), Image co-saliency detection by propagating superpixel affinities, *in* 'Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing', IEEE, pp. 2114–2118.

Tang, M., Gorelick, L., Veksler, O. & Boykov, Y. (2013), Grabcut in one cut, *in* 'Proceedings of the IEEE International Conference on Computer Vision', pp. 1769–1776.

Tompson, J., Goroshin, R., Jain, A., LeCun, Y. & Bregler, C. (2015), Efficient object localization using convolutional networks, *in* 'Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition', pp. 648–656.

Toshev, A., Shi, J. & Daniilidis, K. (2007), Image matching via saliency region correspondences, *in* 'Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition', IEEE, pp. 1–8.

Tsai, C.-C., Li, W., Hsu, K.-J., Qian, X. & Lin, Y.-Y. (2018), 'Image co-saliency detection and co-segmentation via progressive joint optimization', *IEEE Transactions on Image Processing* **28**(1), 56–71.

Tsai, C.-C., Li, W., Hsu, K.-J., Qian, X. & Lin, Y.-Y. (2019), 'Image co-saliency detection and co-segmentation via progressive joint optimization', *IEEE Transactions on Image Processing* **28**(1), 56–71.

Uijlings, J. R., Van De Sande, K. E., Gevers, T. & Smeulders, A. W. (2013), 'Selective search for object recognition', *International Journal of Computer Vision* **104**(2), 154–171.

Vedaldi, A. & Fulkerson, B. (2010), Vlfeat: An open and portable library of computer vision algorithms, *in* 'Proceedings of the ACM International Conference on Multimedia', ACM, pp. 1469–1472.

Vicente, S., Rother, C. & Kolmogorov, V. (2011), Object cosegmentation, *in* 'Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition', IEEE, pp. 2217–2224.

Vidal, R., Ma, Y. & Sastry, S. (2005), 'Generalized principal component analysis (gpca)', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **27**(12), 1945–1959.

Vo, H. V., Bach, F., Cho, M., Han, K., LeCun, Y., Pérez, P. & Ponce, J. (2019), Unsupervised image matching and object discovery as optimization, *in* 'Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition', pp. 8287–8296.

Wang, F., Jiang, M., Qian, C., Yang, S., Li, C., Zhang, H., Wang, X. & Tang, X. (2017), Residual attention network for image classification, *in* 'Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition', pp. 3156–3164.

Wang, J., Yang, Y., Mao, J., Huang, Z., Huang, C. & Xu, W. (2016), Cnn-rnn: A unified framework for multi-label image classification, *in* 'Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition', pp. 2285–2294.

Wang, L., Wang, L., Lu, H., Zhang, P. & Ruan, X. (2016), Saliency detection with recurrent fully convolutional networks, *in* 'Proceedings of the European Conference on Computer Vision', Springer, pp. 825–841.

Wang, T., Borji, A., Zhang, L., Zhang, P. & Lu, H. (2017), A stagewise refinement model for detecting salient objects in images, *in* 'Proceedings of the IEEE International Conference on Computer Vision', pp. 4019–4028.

Wang, W., Shen, J., Li, X. & Porikli, F. (2015), 'Robust video object cosegmentation', *IEEE Transactions on Image Processing* **24**(10), 3137–3148.

Wei, L., Zhao, S., Bourahla, O. E. F., Li, X., Wu, F. & Zhuang, Y. (2019), 'Deep group-wise fully convolutional network for co-saliency detection with graph propagation', *IEEE Transactions on Image Processing* **28**(10), 5052–5063.

Wei, X.-S., Zhang, C.-L., Wu, J., Shen, C. & Zhou, Z.-H. (2019), 'Unsupervised object discovery and co-localization by deep descriptor transformation', *Pattern Recognition* **88**, 113–126.

Wu, Z. & Leahy, R. (1993), 'An optimal graph theoretic approach to data clustering: Theory and its application to image segmentation', *IEEE Transactions on Pattern Analysis and Machine Intelligence* (11), 1101–1113.

Xiao, T., Liu, Y., Zhou, B., Jiang, Y. & Sun, J. (2018), Unified perceptual parsing for scene understanding, *in* 'Proceedings of the European Conference on Computer Vision', pp. 418–434.

Xie, S., Girshick, R., Dollár, P., Tu, Z. & He, K. (2017), Aggregated residual transformations for deep neural networks, *in* 'Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition', pp. 1492–1500.

Xie, S. & Tu, Z. (2015), Holistically-nested edge detection, *in* 'Proceedings of the IEEE International Conference on Computer Vision', pp. 1395–1403.

Xue, J., Li, C. & Zheng, N. (2011), 'Proto-object based rate control for JPEG2000: an approach to content-based scalability', *IEEE Transactions on Image Processing* **20**(4), 1177–1184.

Yan, Q., Xu, L., Shi, J. & Jia, J. (2013), Hierarchical saliency detection, *in* 'Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition', pp. 1155–1162.

Ye, L., Liu, Z., Li, J., Zhao, W.-L. & Shen, L. (2015), 'Co-saliency detection via co-salient object discovery and recovery', *IEEE Signal Processing Letters* **22**(11), 2073–2077.

Yu, H. & Qi, X. (2014), Unsupervised cosegmentation based on superpixel matching and fastgrabcut, *in* 'Proceedings of the IEEE International Conference on Multimedia and Expo', IEEE, pp. 1–6.

Yu, Z. & Wong, H.-S. (2007), 'A rule based technique for extraction of visual attention regions based on real-time clustering', *IEEE Transactions on Multimedia* **9**(4), 766–784.

Yuan, Z.-H., Lu, T. & Wu, Y. (2017), Deep-dense conditional random fields for object co-segmentation., *in* 'International Joint Conferences on Artificial Intelligence', pp. 3371–3377.

Yuan, Z., Lu, T. & Shivakumara, P. (2014), A novel topic-level random walk framework for scene image co-segmentation, *in* 'Proceedings of the European Conference on Computer Vision', Springer, pp. 695–709.

Zagoruyko, S. & Komodakis, N. (2016), 'Wide residual networks', *arXiv preprint arXiv:1605.07146* .

Zahn, C. T. (1970), 'Graph theoretical methods for detecting and describing gestalt clusters', *IEEE Transaction on Computer* **20**(SLAC-PUB-0672-REV), 68.

Zhang, D., Han, J., Li, C. & Wang, J. (2015), Co-saliency detection via looking deep and wide, *in* 'Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition', pp. 2994–3002.

Zhang, D., Meng, D. & Han, J. (2016), 'Co-saliency detection via a self-paced multiple-instance learning framework', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **39**(5), 865–878.

Zhang, D., Meng, D., Li, C., Jiang, L., Zhao, Q. & Han, J. (2015), A self-paced multiple-instance learning framework for co-saliency detection, *in* 'Proceedings of the IEEE International Conference on Computer Vision', pp. 594–602.

Zhang, H., Dana, K., Shi, J., Zhang, Z., Wang, X., Tyagi, A. & Agrawal, A. (2018), Context encoding for semantic segmentation, *in* 'Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition', pp. 7151–7160.

Zhang, J., Zhang, T., Dai, Y., Harandi, M. & Hartley, R. (2018), Deep unsupervised saliency detection: A multiple noisy labeling perspective, *in* 'Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition', pp. 9029–9038.

Zhang, L. & Huang, H. (2012), Hierarchical narrative collage for digital photo album, *in* 'Computer graphics forum', Vol. 31, Wiley Online Library, pp. 2173–2181.

Zhang, L., Tong, M. H., Marks, T. K., Shan, H. & Cottrell, G. W. (2008), 'Sun: A bayesian framework for saliency using natural statistics', *Journal of Vision* **8**(7), 32–32.

Zhang, P., Wang, D., Lu, H., Wang, H. & Ruan, X. (2017), Amulet: Aggregating multi-level convolutional features for salient object detection, *in* 'Proceedings of the IEEE International Conference on Computer Vision', pp. 202–211.

Zhang, P., Wang, D., Lu, H., Wang, H. & Yin, B. (2017), Learning uncertain convolutional features for accurate saliency detection, *in* 'Proceedings of the IEEE International Conference on Computer Vision', pp. 212–221.

Zhang, S., Huang, J., Li, H. & Metaxas, D. N. (2012), 'Automatic image annotation and retrieval using group sparsity', *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics* **42**(3), 838–849.

Zhang, Z., Zhang, X., Peng, C., Xue, X. & Sun, J. (2018), Exfuse: Enhancing feature fusion for semantic segmentation, *in* 'Proceedings of the European Conference on Computer Vision', pp. 269–284.

Zhao, H., Shi, J., Qi, X., Wang, X. & Jia, J. (2017), Pyramid scene parsing network, *in* 'Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition', pp. 2881–2890.

Zhao, R., Ouyang, W., Li, H. & Wang, X. (2015), Saliency detection by multi-context deep learning, *in* 'Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition', pp. 1265–1274.

Zhou, B., Hou, X. & Zhang, L. (2010), A phase discrepancy analysis of object motion, *in* 'Proceedings of the Asian Conference on Computer Vision', Springer, pp. 225–238.

Zhu, G., Ji, Y., Jiang, X., Xu, Z. & Liu, C. (2017), Co-saliency detection based on superpixel clustering, *in* 'Proceedings of the International Con-

ference on Knowledge Science, Engineering and Management', Springer, pp. 287–295.

Zhu, H., Cai, J., Zheng, J., Wu, J. & Thalmann, N. (2013), Salient object cutout using google images, *in* 'IEEE International Symposium on Circuits and Systems', IEEE, pp. 905–908.

Zhu, S. C. & Yuille, A. (1996), 'Region competition: Unifying snakes, region growing, and bayes/mdl for multiband image segmentation', *IEEE Transactions on Pattern Analysis and Machine Intelligence* (9), 884–900.

Zhu, W., Liang, S., Wei, Y. & Sun, J. (2014), Saliency optimization from robust background detection, *in* 'Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition', pp. 2814–2821.

Zou, W., Kpalma, K., Liu, Z. & Ronsin, J. (2013), Segmentation driven low-rank matrix recovery for saliency detection, pp. 78.1–78.11.

Zou, W., Liu, Z., Kpalma, K., Ronsin, J., Zhao, Y. & Komodakis, N. (2015), 'Unsupervised joint salient region detection and object segmentation', *IEEE Transactions on Image Processing* **24**(11), 3858–3873.