Deep learning approaches for object co-segmentation and one-shot segmentation

Li, Guankai

2021

Li, G. (2021). Deep learning approaches for object co-segmentation and one-shot segmentation. Master's thesis, Nanyang Technological University, Singapore.
<https://hdl.handle.net/10356/149937>

<https://hdl.handle.net/10356/149937>

<https://doi.org/10.32657/10356/149937>

This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0).

Downloaded on 26 Apr 2023 17:33:36 SGT

**NANYANG
TECHNOLOGICAL
UNIVERSITY**

SINGAPORE

**DEEP LEARNING APPROACHES FOR OBJECT
CO-SEGMENTATION AND ONE-SHOT
SEGMENTATION**

Li Guankai

School of Computer Science and Engineering

A thesis submitted to the Nanyang Technological University in partial
fulfillment of the requirements for the degree of Master of
Engineering(M.Eng)

2021

Statement of Originality

I hereby certify that the work embodied in this thesis is the result of original research, is free of plagiarised materials, and has not been submitted for a higher degree to any other University or Institution.

10/01/2021

.....
Date

Li Guankai

.....
Li Guankai

Supervisor Declaration Statement

I have reviewed the content and presentation style of this thesis and declare it is free of plagiarism and of sufficient grammatical clarity to be examined. To the best of my knowledge, the research and writing are those of the candidate except as acknowledged in the Author Attribution Statement. I confirm that the investigations were conducted in accord with the ethics policies and integrity standards of Nanyang Technological University and that the research data are presented honestly and without prejudice.

10/01/2021

.....
Date

.....
Asst Prof. Lin Guosheng

Authorship Attribution Statement

This thesis **does not** contain any materials from papers published in peer-reviewed journals or from papers accepted at conferences in which I am listed as an author.

10/01/2021

.....
Date

Li Guankai

.....
Li Guankai

Acknowledgements

First I would like to express my gratitude to my supervisor Asst Prof Lin Guosheng who guided me in the field of Semantic Segmentation in the past two years. I learned lots of research methods and experience from his instructions. He was very willing to share his ideas and experience with me, which inspired me a lot in my research work. When I encountered difficulties, it was his huge encouragement and support that helped me solve the problems step by step and move forward.

I also want to thank my senior schoolmate Zhang Chi in NTU. He cared about my research progress and shared me with his ideas when I was worried and anxious about my research. He was also willing to help me and encourage me when I suffered from frustrating moments in my research experience.

I would like to thank my friends and classmates in NTU, including He Su, Cai Jianfei, Luo Xiangzhong, etc. When I came across the difficulties, they always listened to my complaints about the research experience and shared useful suggestions with me. With their help and support, I could relieve my pressure and never felt helpless. It is quite good luck for me to have all these people to support me.

Abstract

Image co-segmentation is an active computer vision task that aims to discover and segment the shared objects given multiple images. Recently, researchers design various learning-based algorithms to handle the co-segmentation task. The main difficulty in this task is how to effectively transfer information between images to infer the common object regions. In this thesis, we present CycleSegNet, an effective and novel approach for the co-segmentation task. Our network design has two key components: a region correspondence module which is the basic operation for exchanging information between local image regions, and a cycle refinement module which utilizes ConvLSTMs to progressively update image embeddings and exchange information in a cycle manner. Experiment results on four popular benchmark datasets — PASCAL VOC dataset, MSRC dataset, Internet dataset, and iCoseg dataset indicate that our proposed approach greatly outperforms the existing networks and achieves new state-of-the-art performance.

In addition to image co-segmentation, we also explore a method to solve one-shot segmentation with only weak supervision (bounding box). One-shot semantic segmentation has recently gained attention for its strong generalization ability to segment unseen-class images given only limited annotated image. However, existing methods in one-shot object segmentation have mainly relied on manually pixel-wise labeled segmentation masks. The main challenge in this task is limited data and weak supervision. In this thesis, we present an effective approach, which utilizes the recent weakly-supervised semantic segmentation method to generate pseudo mask labels in the bounding box regions and then integrates the detailed information and correlation between support image and query image for solving one-shot image segmentation. Extensive experiments on the PASCAL-5ⁱ dataset show that our weakly-supervised method narrows down the performance gap between bounding box supervision and pixel-wise annotations, and performs comparably with the state-of-the-art fully-supervised one-shot methods.

List of Figures

1.1	An illustration of our motivation on object co-segmentation. Previous methods exchange image-level semantics as cues for image co-segmentation (up), while our method establishes correspondence between local image regions and refines the predictions iteratively (down).	2
1.2	Illustration of our proposed weakly-supervised one-shot segmentation. Compared with the previous one-shot image segmentation, we first utilize a weakly-supervised semantic segmentation network to generate the pseudo mask of support image instead of requiring pixel-level annotation. Then our one-shot segmentation model will take the support set and the query image as inputs to generate the prediction for query image.	3
3.1	Our network for image co-segmentation with paired or group input images. The main components in our network include a Cycle Refinement Module (CRM) for feature updating and a Region Correspondence Module(RCM) for information exchanging.	20
3.2	Architecture of Region Correspondence Module. $\oplus$ indicates element-wise add and $\otimes$ indicates matrix multiplication.	21
3.3	Our model variant utilizing multi-level features in the encoder for image co-segmentation. We apply the CRM module (described in Section. 3.2.2 and illustrated in Figure. 3.1) to features from different layers independently and fuse the resulting features using the Channel Attention module (CAM) proposed in [66] during the upsampling stage.	21
3.4	Predictions from different refinement steps. We decode the representations in each time step and the result shows that our proposed cycle refinement module can consistently improve the prediction.	22
3.5	Qualitative results(Jaccard) of our network on the PASCAL VOC, MSRC, Internet, iCoseg dataset.	22
4.1	The whole architecture of our one-shot segmentation network. DEM indicates Detail Extraction Module and FFM represents Feature fusion Module. Given a query image and a support image, we first employ a weight-shared CNN to extract their features in different levels. Then we utilize the graph-based method proposed in [72] to model the correspondence between the middle-level features from support image and query image, and our proposed Detail Extraction Module to capture object detail information based on low-level feature map. After that, the detailed information and correspondence information are processed by Feature Fusion Module and further upsampled to generate the predicted mask.	26
4.2	Architecture of SEAM [62]	28

4.3	Architecture of our proposed Detail Extraction Module.	29
4.4	Architecture of our proposed Feature Fusion Module(FFM).	31
4.5	Qualitative results of our Pseudo Masks Generation. The first, third, and fifth columns are the input images. The second, fourth, and sixth columns are our generated pseudo masks.	32
4.6	Qualitative results of our weakly-supervised one-shot segmentation network. The first, third, and fifth columns are the input images. The second, fourth, and sixth columns are our predicted masks.	33

List of Tables

3.1	Ablative experiments of the proposed model. “ M_{cat} ”, “ M_{mul} ” and “ M_{cross} ” denote three kinds of ways to exchange information between images. “ M_{cycle} ” denotes our proposed cycle refinement module. Please refer to Section. 3.3.3 for the descriptions of the baseline models.	14
3.2	Our network with different refinement steps N . With the refinement step increasing, the performance can increase consistently.	14
3.3	Our network performance with regional representations of different sizes in the cross region module.	15
3.4	Performance of our method with features from different stages in the encoder network. We test ResNet-34 and VGG16 as the network encoders. Using multi-level features can effectively boost network performance. . .	15
3.5	Performance of our network with Lovász-Softmax loss function and the standard cross entropy loss function.	15
3.6	The results(Jaccard) of group segmentation with different sampling strategies and different numbers of input images on the Internet dataset. Our network can handle more than two images at a time, which generate better results than our network with paired input images.	16
3.7	Comparison with the state-of-the-art performance of pairwise co-segmentation and group segmentation tasks on four benchmark datasets: Pascal VOC , MSRC , Internet and iCoseg . Our model achieves new state-of-the-art performance on all datasets.	23
4.1	Ablation experiments on the Pseudo Masks Generation process. Ours-origin indicates using bounding box labels as supervision in one-shot training, and Ours-pseudo indicates utilizing generated pseudo labels to supervise one-shot training. Both of the methods use our proposed one-shot segmentation network.	29
4.2	Ablation experiments on the Detail Extraction Module(DEM). Ours-init indicates discarding DEM in our proposed model, Ours-Conv means using a convolution layer with kernel size 3x3 to replace DEM, and Ours-DEM represents our proposed model.	29
4.3	The results of different fusion operations on the PASCAL-5 ⁱ dataset. Ours-sum, Ours-concat, Ours-FFM represent using element-wise summation, concatenation and our proposed FFM to fuse feature in different levels respectively.	30
4.4	Ablation experiments on the Muti-Scale Evaluation. All predictions are rescaled to the same image size and merged by average.	30

4.5	mean IoU of one-shot segmentation on the PASCAL-5 ⁱ dataset under pixel-wise supervision and bounding box supervision. Ours-MS represents our proposed method with Multi-Scale input during testing	30
-----	--	----

Contents

1	Introduction	1
1.1	Problem Background	1
1.2	Major Contributions	3
1.3	Outline of the Thesis	4
2	Literature Review	5
2.1	Object Co-segmentaion	5
2.2	Iterative Refinement in Segmentation	5
2.3	Co-saliency Detection	6
2.4	Few-shot Semantic Segmentation	6
2.5	Weakly-supervised Semantic Segmentation	7
3	CycleSegNet: Object Co-segmentation with Region Correspondence and Cycle Refinement	8
3.1	Introduction	8
3.2	Methodology	10
3.2.1	Cycle Refinement Module	10
3.2.2	Region Correspondence Module	11
3.2.3	Group Segmentation	12
3.2.4	Multi-Level Features for Co-segmentation	13
3.3	Experiment Results	13
3.3.1	Implementation Details.	13
3.3.2	Datasets and Evaluation Metric	15
3.3.3	Ablation Study	16
3.3.4	Comparison with the State-of-the-arts	18
3.4	Conclusion	18
4	Weakly-supervised One-shot Segmentation	24
4.1	Introduction	24
4.2	Methodology	25
4.2.1	Problem Definition	26
4.2.2	Pseudo Masks Generation	27
4.2.3	One-shot Segmentation Network	27
4.3	Experiment Results	29
4.3.1	Implementation Details	29
4.3.2	Datasets and Evaluation Metrics.	31
4.3.3	Ablation Study	31
4.3.4	Comparison to the Pixel-wise Annotation	34

4.4	Conclusion	34
5	Conclusion and Future Work	36
5.1	Conclusion	36
5.2	Future Work	37

Chapter 1

Introduction

1.1 Problem Background

Image segmentation is one of the basic computer vision problems which intends to segment images into semantically different regions. Rother et al. [46] firstly proposed the task of object co-segmentation, aiming to discover and segment the shared objects from a set of images, which has been used in various applications like image retrieval [56], 3D reconstruction [40], image similarity measures [46], image matching [8, 71] and video object tracking and segmentation [46, 36]. In recent years, the booming development of deep neural networks [37, 2, 9] promotes the remarkable success in the field of image segmentation. However, these models can not be directly applied to the co-segmentation tasks, as the outputs are conditioned on the pairwise or group-wise relations between input images.

Many existing CNN based co-segmentation methods [33, 7, 3] often apply a pair of parameter-shared Siamese network to generate feature representations of two images and use various methods to transfer information between them to make predictions. The essence of these designs is that two images exchange information about what objects are existing on both sides, and networks refer to the exchanged information to make the pixel-level predictions of the common object regions.

The biggest challenge here is how to effectively transfer useful information via deep neural networks. Directly exchanging the image representations is unrealistic, because images have structured representations and no correspondence information between elements is provided. Therefore, we can not simply multiply or concatenate them when the elements are not aligned. Many recent works [3, 7] try to bypass the correspondence problem by squeezing the image structures and exchange information in the form of a global image representation which is usually achieved by the global average pooling. As a result, there is no correspondence problem and the prediction of each pixel location can refer to the global vector. However, squeezing the image structures into a global representation inevitably loses local discriminative information which can be useful to locate common objects from the scene. Moreover, as the image content often has a complex composition, including the cluttered background and objects from different classes, the image-level representation may also introduce noise into the cues for conditional predictions. In this thesis, we propose to solve the aforementioned problems from two perspectives, which are illustrated in Fig. 1.1.

Although the semantic segmentation models gain great success to segment the objects after training on a large scale of annotated data, they lack the ability to segment object

Figure 1.1: An illustration of our motivation on object co-segmentation. Previous methods exchange image-level semantics as cues for image co-segmentation (up), while our method establishes correspondence between local image regions and refines the predictions iteratively (down).

from new categories without the additional finetuning. To address this challenge, there emerges a new research topic which learns new categories information from limited annotated data, known as few-shot image segmentation. Few-shot image segmentation, which aims to predict unseen categories with a few annotated images after training, is another challenging and critical segmentation task and gains increasing popularity. Compared with the common segmentation approaches, which provide masks as results for given input images, few-shot segmentation focuses on mining the correlation between the query and the support image, which greatly improves the generalization ability given the limited annotations. Researchers have already investigated this topic in [50], [70], [72], [59].

However, existing methods in few-shot object segmentation have mainly depended on high-cost pixel-wise segmentation masks. Limited research has been conducted on using weak supervision in the training process of few-shot segmentation. This is our key setting (one shot & bounding boxes level supervision) that our thesis attempts to explore. The main difficulties in this task are limited data and weak supervision, assuming pixel-wise annotations are not provided in the training process. The set-up is very practical and common in the real world, since the ability to segment objects after seeing one or a few reference images is a natural case for humans.

Figure 1.2: Illustration of our proposed weakly-supervised one-shot segmentation. Compared with the previous one-shot image segmentation, we first utilize a weakly-supervised semantic segmentation network to generate the pseudo mask of support image instead of requiring pixel-level annotation. Then our one-shot segmentation model will take the support set and the query image as inputs to generate the prediction for query image.

In recent time, many researches are devoted to weakly supervised semantic segmentation, which only require weak supervisions, such as scribbles, bounding boxes, or even image-level classification labels, aiming to reduce the performance gap between full and weak supervisions. Most of the advanced weakly supervised semantic segmentation networks exploit the class activation map(CAM) [83], which has been proven to discover and extract the local discriminative parts of the target object. The discriminative regions highlighted by CAMs will be used as seed and expanded to the whole object area. Since in the one-shot or few shot cases, the testing classes and the training classes are non-overlapping, thus these methods are not applicable directly. However, we can utilize these weakly-supervised methods in the bounding box regions to generate pseudo masks for our one-shot training and provide more accurate object information. It is still an open problem on how to alleviate the pixel-wise labeling requirements in few-shot segmentation. In this thesis, we present our solution to overcome the aforementioned problems problem. Fig. 1.2 shows the overview of our proposed method.

1.2 Major Contributions

Our main contributions of this thesis can be stated as follows:

Firstly, to solve the task of object co-segmentation, we design a novel network named CycleSegNet with region correspondence module and cycle refinement module. The region correspondence module is proposed to maintain local discriminative representations

and exchange information between paired images. The cycle refinement module employs ConvLSTMs to progressively refine both the exchanged representations and network predictions. In addition to inputs of paired images, our network can also handle inputs of grouped images with the same network and parameters. The evaluation of four popular benchmarks — Pascal VOC dataset, MSRC dataset, Internet dataset, and iCoseg dataset, demonstrates that our proposed method greatly outperforms the existing co-segmentation models.

Secondly, to alleviate the labeling requirements, we explore an effective low-cost attempt in the setup of one-shot segmentation to support weak (bounding box label) supervision. As the bounding box labels are not as accurate as the pixel-wise annotations, we reference the recent weakly supervised segmentation approach to generate pseudo labels, which are used to supervise the training process of one-shot segmentation. The proposed Detail Extraction Module can extract detailed information at different levels and our designed Feature Fusion Module is capable to integrate the detail information and correspondence information between the query image and the support image. We extensively evaluate our proposed method on the PASCAL-5ⁱ dataset and our weakly-supervised approach(53.1%) can perform close to the current state-of-the-art fully supervised one-shot segmentation models(56.0%).

1.3 Outline of the Thesis

The remaining parts of this thesis are organized as follows. Chapter 2 reviews the related works, including the object co-segmentation, iterative refinement in segmentation, co-saliency detection, few-shot semantic segmentation, and weakly supervised semantic segmentation. Chapter 3 presents our work of Object Co-segmentation with Region Correspondence and Cycle Refinement (CycleSegNet). It proposes a region correspondence module and cycle refinement module to exchange information between local image regions in a cycle manner, which aim to transfer information between images and segment the shared objects in multiple images. Chapter 4 presents our work of weakly-supervised one-shot segmentation. It firstly introduces a weakly-supervised semantic segmentation approach to obtain pseudo masks in the bounding boxes, and then proposes Detail Extraction Module and Feature Fusion Module to integrate the detail information and feature correspondence to solve the task of one-shot segmentation in weak supervision. The thesis is concluded in Chapter 5 and the potential future works are discussed.

Chapter 2

Literature Review

2.1 Object Co-segmentation

Early work on addressing object co-segmentation generally compares the visual features of paired images, such as foreground color histogram [46], SIFT [39], saliency [6]. Besides, [29] indicates that finding discriminative features can be well adapted to solve the object co-segmentation. The method in [48] applies region matching method to establish correspondences between the common objects. Rubio *et al.* [48] trains SVM classifier based on a Gaussian Mixture Model to capture the correspondence between different regions of input images. Rubinstein *et al.* [47] utilize dense correspondences and visual saliency to find out the sparsity and visual variability of the shared objects throughout the database. In the past few years, some CNN-based networks have been devoted in the object co-segmentation task. For instance, Quan *et al.* [41] propose a manifold ranking algorithm by combining the features obtained from VGG network [54] and handcrafted features for superpixels to implement object co-segmentation. [33] is the first one to utilize a Fully Convolutional Network [37] in the setup of object co-segmentation. The proposed method utilizes a Siamese encoder-decoder architecture and a mutual correlation layer to segment the co-existing objects in multiple images. Chen *et al.* [7] leverage the attention mechanism for this setup. They propose a channel-wise learner to exchange feature information and enhance the common semantic information. Hsu *et al.* [23] present an end-to-end method using neural networks which dealt with this task in an unsupervised manner. Zhang *et al.* [77] design a spatial modulator for group-wise mask learning and a semantic modulator for co-category classification. Li *et al.* [31] propose a recurrent module to generate the group representation containing the synergetic information among multiple relevant images, which was broadcasted to each individual image to make the co-segmentation predictions.

2.2 Iterative Refinement in Segmentation

Many previous works have explored using interactive structure to optimize segmentation results. For example, McIntosh *et al.* [38] adopt convolutional LSTM [22] to control computation budgets via a number of iteration steps. Wang *et al.* [61] propose to use saliency maps to iteratively refine weakly segmentation results. Lin *et al.* [34] propose to use graphical models to optimize the mask learned by scribble supervision for semantic segmentation. In [24] and [70], iterative structures are also used to improve few-shot

segmentation results. Compared with previous iterative structures, the main difference in our design is that our two branches can optimize each other and meanwhile refine their own predictions in a cycle manner.

2.3 Co-saliency Detection

Co-saliency detection [63, 74, 15, 17, 73, 75] is a closely related topic which aims to identify the common and salient objects from a group of images. Zhang *et al.* [74] explore intrasaliency prior transfer and deep intersaliency mining for co-saliency detection. Wei *et al.* [63] introduce the group-wise feature representation learning and the collaborative learning to address the co-salient object discovery problem. To enable fast common information learning, Fan *et al.* [15] propose a network to simultaneously embed the appearance and semantic features through a co-attention projection strategy. Zhang *et al.* [79] utilize multi-level CNN features to improve the RGB-T salient object detection, which shares similar spirits with our model variants. Li *et al.* [32] employ attention mechanisms to integrate cross-modal and cross-level complementarity from multi-modal data for RGB-D salient object detection. We also utilize attention mechanisms in our design, but the attention is used to establish regional connections between images, which has a different purpose. Zhang *et al.* [78] design a feature fusion network to undertake the RGB-T saliency detection task, which includes multi-scale, multi-modality, and multi-level feature fusion modules. Compared with co-saliency detection, object co-segmentation must identify objects from multiple categories from a complex scene, which are not necessarily the salient areas in images. For example, the buses and the people in Fig. 1.1 are both common objects in the images.

2.4 Few-shot Semantic Segmentation

Few-shot semantic segmentation has recently gained increasing popularity due to its strong generalization ability to predict unseen-class images with only limited annotated samples. Shaban *et al.* [50] propose the first few-shot segmentation method using a dual branched neural network to process support and query images separately. Weight hashing operation is designed to integrate semantic context and generate the parameters for the query branch. Zhang *et al.* [70] take inspiration from metric learning and propose a dense comparison module with a siamese-like architecture to compute the similarity by simply concatenating query and support features, and an iterative optimization to gradually refine the results. Wang *et al.* [59] introduce an alignment regularization on learned prototype representations. It firstly aligns the query features to the embedding space of support images and also predicts the support masks based on the prototype embedding of query features. Zhang *et al.* [72] utilize attention mechanism to construct multi-level region-based connections between support image and query image to segment the target objects. However, most of existing few-shot segmentation methods require training with a large number of pixel-wise annotated examples in some particular classes. And if the models need to adapt to unseen classes, they still need some image-mask pairs to feed the model as guidance.

2.5 Weakly-supervised Semantic Segmentation

Weakly-supervised semantic segmentation aims to explore a feasible method for semantic segmentation to reduce the cost of pixel-level annotation. Wei *et al.* [64] propose a two-stage architecture to generate segmentation masks. First, the coarse parts of the object are determined by the CAM[83]. The segmentation part, such as DeepLab, is then trained to produce the segmentation masks. Zhang *et al.* [80] take inspiration from the CAM and propose to use classification network to discover the regions of objects. Based on the original CAM, [1] learns the semantic affinity metric within a local area without additional supervision, which is combined with the random walk to generate the segmentation masks. Wang *et al.* [62] discover the output CAMs of rescaled input images are not consistent, thus introduce a regularization into the weakly-supervised segmentation network. They also utilize the self-attention mechanism to integrate more context information before the network output. To address the problem of weakly-supervised semantic segmentation, Sun *et al.* [55] propose two kinds of co-attention to capture the cross-image semantic relations and differences for comprehensive object pattern mining.

Chapter 3

CycleSegNet: Object Co-segmentation with Region Correspondence and Cycle Refinement

3.1 Introduction

Image co-segmentation is an active computer vision topic with a long research history, aiming to discover and segment the common objects jointly from multiple images. Image co-segmentation algorithms have shown their usages in a number of computer vision tasks, including image retrieval [56], 3D reconstruction [40], photo collections [46], image matching [8, 71] and video object tracking and segmentation [46, 36]. Recently, data-driven deep neural networks based methods attract wide interest in the literature. The powerful Deep Neural Networks along with the challenging evaluation benchmarks built upon large-scale public datasets have brought this task to a new era and make it more challenging. Although deep neural networks have shown remarkable success in many other segmentation tasks, such as semantic segmentation [37, 16, 28, 72, 70], interactive segmentation [49], and instance segmentation [21, 76, 67], their models can not be directly applied to the co-segmentation tasks, as the outputs are conditioned on the pairwise or group-wise relations between input images.

Many existing CNN based methods [33, 7, 3] often apply a pair of parameter-shared Siamese network to generate feature representations of two images and use various methods to transfer information between them to make predictions. The essence of these designs is that two images exchange information about what objects are existing on both sides, and networks refer to the exchanged information to make the pixel-level predictions of the common object regions. The biggest challenge here is how to effectively transfer useful information via deep neural networks. Directly exchanging the image representations is unrealistic, because images have structured representations and no correspondence information between elements is provided. Therefore, we can not simply multiply or concatenate them when the elements are not aligned. Many recent works [3, 7] try to bypass the correspondence problem by squeezing the image structures and exchange information in the form of a global image representation which is usually achieved by the global average pooling. As a result, there is no correspondence problem and the prediction of each pixel location can refer to the global vector. However, squeezing the image

structures inevitably loses information and introduces noise, as the image content may have a complex composition, such as a cluttered background and objects from different classes. Directly mixing the pixel-level representations makes the useful information unclear and definitely raises the difficulty when using such information as the reference to locate common objects.

In this paper, we propose to solve the problems above from two perspectives. First, we argue that a desired algorithm should maintain the image structures and let the exchanged information contain local discriminative representations. To establish the correspondence between local representations, we design a region correspondence module that utilizes the attention mechanism to capture the correlations between local regional representation. Each pixel location can find the most relevant regions in the other image and the prediction of each pixel refers more to the relevant regions based on the attention value. We replace the global average pooling operation with our proposed region correspondence module to exchange information between two branches.

Second, we design a cycle refinement module to refine feature representations and predictions in a cycle manner. Our motivation is based on an observation that a good feature representation that contains more information about the common classes and less about cluttered background can provide more accurate information when it is transferred to the other branch as a guidance. At the same time, a better prediction can further help pinpoint the region of the target class and suppress the noise from the background thus producing better representations to transfer. Therefore we can improve both the quality of the exchanged feature representation and the network prediction progressively. Based on such intuition, we propose to optimize the co-segmentation prediction and the feature representations iteratively. We employ a pair of parameter-shared convolutional LSTMs (ConvLSTMs) to achieve this goal, which is operated on the bottleneck of an encoder-decoder network. At each time step, the LSTM cell takes the exchanged embedding from the opposite branch as the input and update its own embedding, which is then sent to the opposite branch as the exchanged information in the next step. The cycle refinement module can thus continuously exchange information between two images and refine the image representations.

An advantage of our design is that our network for co-segmentation can handle the input with paired images or a group of images with the same model and network parameters. This is because the key operation in our network design is based on attention mechanism and does not require a fixed number of pixels on each side. We can transform the network-like relations between images to a 1-versus-others relation to handle the group segmentation task. In the network for group segmentation, all images are encoded by the same encoder and their representations are updated by extracting information from all other images with the cycle refinement modules.

Finally, we also develop a model variant that exploits multi-level features to better segment the common objects in two images. As is often observed in the CNN visualization literature [69], high-level features contain more semantic information while middle-level features correspond to object parts that may be shared across different object categories. Therefore, it is useful to establish region connections at multiple semantic levels for predictions. We apply our proposed modules to the features at different stages independently and then fuse their results during the upsampling process. As a result, our network can better locate the pixels belonging to the common objects.

To validate our network designs, we implement various experiments on multiple datasets to analyze each component in our design. We empirically demonstrate that

our design can significantly improve the performance over existing models. Main contributions of this work are summarized as follows:

- We propose a region correspondence module as a basic operation to exchange information between paired images for the co-segmentation task. The proposed module maintains local discriminative representations and utilizes the attention mechanism to directly establish correspondence between image regions.
- We propose a cycle refinement module that employs ConvLSTMs to progressively refine both the exchanged representations and network predictions.
- We demonstrate that our model can handle inputs of paired images and grouped images with the same network and parameters.
- We develop a model variant that exploits multi-level features to undertake co-segmentation. The multi-level version of our network can better locate common objects and boost network performance.
- Experiments on four popular benchmarks — Pascal VOC dataset, MSRC dataset, Internet dataset and iCoseg dataset, demonstrate that our proposed network on both co-segmentation and group segmentation tasks greatly outperforms the existing approaches and achieves new state-of-the-art performance.

3.2 Methodology

In this section, we present our framework for image co-segmentation task. We begin with our network description in the case of paired input images. Each image is encoded and decoded with a parameter-shared Siamese network. The network design has an encoder-decoder structure, and the two branches exchange information at the network bottlenecks. We first describe our model version that only exploits the features at the last layer of the encoder for information exchange, which is shown in Figure. 3.1. Then, we describe the model variant that uses multi-level features. Finally, we describe how to extend our network to the group-segmentation task without learning new parameters.

3.2.1 Cycle Refinement Module

The cycle refinement module (CRM) aims to update the image embeddings by incorporating information from the other image. The overall structure of the cycle refinement module can be found in Figure. 3.1. In each step, the CRM implicitly compares the co-occurrent embeddings of two images and lets the representations focus on common objects progressively. To achieve this goal, we employ a ConvLSTM at the network bottleneck to undertake the tasks of information exchanging and representation updating. The LSTM has proven its success in numerous computer vision and NLP tasks. Its feedback connections and memory cells make it well-suited for tasks with sequential inputs. The ConvLSTM further replaces the linear transformations in LSTM with convolutional kernels, such that the LSTM owns large field-of-views and is suited for process

2-dimensional data. A typical ConvLSTM cell has the following structures:

$$i_t = \sigma(W_{xi} * X_t + W_{hi} * H_{t-1} + b_i), \quad (3.1)$$

$$f_t = \sigma(W_{xf} * X_t + W_{hf} * H_{t-1} + b_f), \quad (3.2)$$

$$o_t = \sigma(W_{xo} * X_t + W_{ho} * H_{t-1} + b_o), \quad (3.3)$$

$$\tilde{C}_t = i_t \odot \tanh(W_{xc} * X_t + W_{hc} * H_{t-1} + b_c) \quad (3.4)$$

$$C_t = f_t \odot C_{t-1} + i_t \odot \tilde{C}_t, \quad (3.5)$$

$$H_t = o_t \odot \tanh(C_t), \quad (3.6)$$

where i_t is the input gate, which controls the activation of the new input information; f_t is the forget gate that clears the past cell status; o_t is the output gate, controlling whether the latest cell output C_t will be propagated to the final state H_t ; W is convolutional kernels; $*$ donates the convolution operator and $\odot$ donates the Hadamard product. At each time step t , the ConvLSTM cell takes in an input X_t and updates the hidden states H_t and cell state C_t , which are both initialized as the image representations generated by the encoders. The updated cell state C_t is then sent to the region correspondence module, which compares the cells from both sides and generates the inputs of the next step:

$$X_t^A = \text{cross_region}(C_{t-1}^A, C_{t-1}^B), \quad (3.7)$$

$$X_t^B = \text{cross_region}(C_{t-1}^B, C_{t-1}^A). \quad (3.8)$$

At the last step, the hidden state H is decoded by the decoder and generates the prediction mask of the common objects. Based on this recurrent process, both the quality of the transferred representations and the accuracy of predictions can be improved iteratively. The number of steps for information exchanging is flexible. In our experiment, we investigate the influence of the step number on the performance and the computation cost it takes. Our proposed CRM for image co-segmentation has a symmetric structure design and all the operations on both sides are parameter-shared.

3.2.2 Region Correspondence Module

As we have seen in the cycle refinement module, the key component that exchanges information between two images is the region correspondence module. Previous works often achieve this goal through a global representation as the exchanged information—the exchanged global representation is then fused with the dense image representations by multiplication or concatenation. However, as image segmentation is a dense prediction task, pixel-wise local discriminative representations are crucial information to inference the common object regions. Therefore, we maintain the local representations of images and use attention mechanism to establish the regional connections between two images. Specifically, the input to the region correspondence module is the current representations of two images $C_t^A \in R^{H_A \times W_A \times C}$ and $C_t^B \in R^{H_B \times W_B \times C}$ generated by the LSTMs, where H , W and C denotes the height, width and channel number of the feature map. When we want to transfer information from image I_B to image I_A , we first get the regional representations of image I_B by applying the ROI pooling to the representation C_t^B which

downsamples the image representations to a fixed spatial size $\hat{H} \times \hat{W}$:

$$C_t'^B = \text{Conv}(\text{ROI}_{avg}(C_t^B) || \text{ROI}_{max}(C_t^B)) \quad (3.9)$$

The regional representations can provide context information of local regions, which are proven useful in many previous segmentation works [82]. We apply both the ROI average pooling and the ROI max pooling to C_t^B , and fuse their results by convolutions to get $C_t'^B$. The operations here share some similarities with the Channel Attention Block proposed in [65], where the global average pooling and global max pooling are used together to generate channel attentions. Then we compute the similarity between each feature point in C_t^A and $C_t'^B$ by dot product, which is implemented in parallel by matrix multiplication:

$$S_t^{AB} = W^A(C_t^A) \times (W^B(C_t'^B))^T, \quad (3.10)$$

where $S_t^{AB} \in R^{H^A W^A \times \hat{H} \hat{B}}$, and W is the linear transformation function followed by ReLU non-linearity, implemented as 1×1 convolution. For notational convenience, we omit the reshape operations in Equation.3.10. Intuitively, the function W encodes the feature representations into a space for computing feature similarity and the affinity matrix S_t^{AB} can reflect the correlations between local representations. Then, we normalize the affinity matrix by softmax and use the normalized affinity matrix to query features from the regional representations of image I^B :

$$X_t^{AB} = \text{softmax}(S_t^{AB}) \times C_t'^B, \text{ where } X_t^{AB} \in R^{H_A W_A \times C} \quad (3.11)$$

We also incorporate the global statistics of image I^B into the exchanged information by global average pooling, and upsample it to the same spatial size of image I^A . The final output of region correspondence module is

$$\text{cross_region}(C_t^A, C_t^B) = (X_t^{AB} + \text{Up}(\text{AvgPool}(C_t^B)))/2. \quad (3.12)$$

When transferring information from image I^A to image I^B , we can simply reverse the notation of A and B in the above equations, and apply the same operation without learning new parameter. As we can see above, the exchanged information for each pixel position is different. Every pixel can attend all the local regions in the other image and selectively extract information from different regions.

3.2.3 Group Segmentation

Group segmentation is a special case of the co-segmentation task, where the goal is to discover and segment the common objects in a group of images. An advantage of our network is that the image co-segmentation model for paired input images can also be applied to the group segmentation task with only minor modifications and without learning new parameters. Since all the operations in the region correspondence module do not require a fixed spatial size of the representations, we can use the region correspondence module to directly transfer information from all other images to the target image by treating all other images as an big virtual image. Figure. 3.2 illustrates how our co-segmentation network can be used to handle the group segmentation task. For example, when a group segmentation task has three input images, the prediction of an image should refer to information from the other two images. We achieve this goal by constructing an affinity matrix that includes the similarity between the target image and all other

images. We can therefore extract information from all other image based on the attention distribution. Specifically, the Equation. 3.9 and Equation. 3.10 become:

$$C_t'^B = Conv(ROI_{avg}(C_t^B) || ROI_{max}(C_t^B)), \quad (3.13)$$

$$C_t'^C = Conv(ROI_{avg}(C_t^C) || ROI_{max}(C_t^C)), \quad (3.14)$$

$$C_t'^{BC} = Flatten(C_t'^B) || Flatten(C_t'^C), \quad (3.15)$$

$$S_t^{A-BC} = W^A(C_t^A) \times (W^B(C_t'^{BC}))^T, \quad (3.16)$$

where Flatten is the operation that reshapes the 2D tensor to 1D tensor and $||$ indicates the concatenation operated along the spatial dimension. To make it more clear, the shapes of the tensors above are listed here: $C_t'^{BC} \in R^{(\hat{H}\hat{W}+\hat{H}\hat{W}) \times C}$ and $S_t^{A-BC} \in R^{(H^A W^A \times (\hat{H}\hat{W}+\hat{H}\hat{W}))}$. Thus, we can reuse all other operations in the region correspondence module to undertake the group segmentation task. The above operations transform the network-like relations between individual images to a 1-versus-others relation, such that we can use our co-segmentation model to solve the group segmentation problem. We can repeat such operations to make predictions for each of the images in the group using the same parameters.

3.2.4 Multi-Level Features for Co-segmentation

As the objective of co-segmentation is to segment common objects in two images, the similar regions in two images are important cues. We observe that such similarity may exist on multiple semantic levels. For example, if both the training and testing images contain the class *car*, high-level features that correspond to object categories are useful for locating such category. However, the testing images also contain novel classes that are never seen in the training process. In this case, middle-level features that correspond to object parts are useful, as they are more likely to be shared across classes. Based on such intuition, we design a model variant that exploits multi-level features in the encoder to undertake the co-segmentation task. We apply our cycle-refinement module to features on different levels individually, and fuse their representations in the upsampling stage, as shown in Figure. 3.3. We adopt the channel attention module proposed in [66] to fuse the representations from different levels. The multi-level version network can also apply to the group-segmentation task. Multi-level features have already been proven useful in various vision tasks, including semantic segmentation [45] and object detection [44]. Our experiments in Section.4.3 show that using multi-level features to reason the common object region is also very effective in the co-segmentation task.

3.3 Experiment Results

3.3.1 Implementation Details.

Network. The best performance of our network is achieved when we leverage the ResNet34 [20] as the encoder network and use multi-level features from the last three layers; the number of iterative steps is set as 7; the ROI region size is set as 2×2 . If not further emphasized, we use them as the default setup in the following experiments. The encoder is pretrained on ImageNet dataset [11]. The decoder in our model version

M_{cat}	M_{mul}	M_{cross}	M_{cycle}	$\mathcal{P}$ (%)	$\mathcal{J}$ (%)
✓				77.2	58.6
				81.3	62.9
				83.6	64.2
		✓		87.1	68.2
	✓	✓	✓	91.8	70.4

Table 3.1: Ablative experiments of the proposed model. “ M_{cat} ”, “ M_{mul} ” and “ M_{cross} ” denote three kinds of ways to exchange information between images. “ M_{cycle} ” denotes our proposed cycle refinement module. Please refer to Section. 3.3.3 for the descriptions of the baseline models.

iterations(N)	$\mathcal{P}$ (%)	$\mathcal{J}$ (%)	Time(s)
$N=2$	92.2	71.9	0.085
$N=3$	93.8	73.1	0.088
$N=4$	94.2	74.0	0.096
$N=5$	94.8	74.5	0.105
$N=6$	95.3	75.2	0.109
$N=7$	95.8	75.4	0.114

Table 3.2: Our network with different refinement steps N . With the refinement step increasing, the performance can increase consistently.

with single-level feature encoder is identical with the decoder structure used in [7], which contains five blocks of bilinear upsampling layers and convolutional layers.

Training. We use Adam algorithm [30] to optimize our whole network in an end-to-end manner. The learning rate is set to 1e-5 and weight decay is 0.0005. As previous models use different training datasets in this task, we adopt corresponding training datasets for fair comparison. Following [33, 7], we generate our training data from the training set of Pascal VOC2012 dataset while following [31, 77], we generate our training data from COCO dataset [35]. For both training and evaluation, we resize every image from the datasets to the spatial size of 512×512 . We choose the batch size of 4 and train our network on the NVIDIA 1080TI GPU for about 40K iterations. We use Lovász-Softmax proposed in [5] as the loss function, as we find it can yield slightly better performance than the pixel-wise cross-entropy loss. The complete loss function is given as:

$$\mathcal{L} = \frac{1}{C} \sum_c \Delta_{J_c}^-(m(c)),$$

$$m_i(c) = \begin{cases} 1 - p_i(c) & \text{if } c = y_i^*, \\ p_i(c) & \text{otherwise,} \end{cases} \quad (3.17)$$

where C indicates the amount of class, and y_i^* and $p_i(c)$ denote the ground truth label and the predicted probability of pixel i for class c , respectively. $\Delta_{J_c}^-$ is the Lovász extension of the Jaccard Index.

region-size	$\mathcal{P}(\%)$	$\mathcal{J}(\%)$
original size	92.9	73.1
2x2	95.8	75.4
3x3	94.0	74.1
4x4	95.1	74.8
5x5	93.6	74.0
6x6	94.8	74.4
7x7	94.6	74.3

Table 3.3: Our network performance with regional representations of different sizes in the cross region module.

res_5	res_4	res_3	$\mathcal{P}(\%)$	$\mathcal{J}(\%)$
✓			91.8	70.4
✓	✓		93.5	72.5
✓	✓	✓	95.8	75.4
vgg_5	vgg_4	vgg_3	$\mathcal{P}(\%)$	$\mathcal{J}(\%)$
✓			91.2	69.8
✓	✓		93.1	71.8
✓	✓	✓	94.9	74.2

Table 3.4: Performance of our method with features from different stages in the encoder network. We test ResNet-34 and VGG16 as the network encoders. Using multi-level features can effectively boost network performance.

loss function	$\mathcal{P}(\%)$	$\mathcal{J}(\%)$
Lovász-Softmax	95.8	75.4
cross entropy	94.1	74.4

Table 3.5: Performance of our network with Lovász-Softmax loss function and the standard cross entropy loss function.

3.3.2 Datasets and Evaluation Metric

To evaluate the performance, we compare our proposed method with existing approaches on four widely-used benchmark datasets for object co-segmentation, including the Pascal VOC dataset [13], MSRC dataset [51], Internet dataset [47] and iCoseg dataset [4].

Pascal VOC. As some of the previous literature adopts PASCAL VOC2012 dataset and some previous work uses PASCAL VOC 2010 dataset, in our thesis we adopt both of these two dataset. PASCAL VOC2012 has 20 foreground object categories with 1464 training samples and 1449 validation samples totally. In the task of object co-segmentation, the validation set is splitted into two sets: validation set (724 images) and test set(725 images) following the previous work [33, 7, 60]. Following the setting in [31, 77], we also adopt the PASCAL VOC 2010 dataset which contains 20 categories with 1037 images in the evaluation stage.

MSRC. Following [33, 77], we adopt the same subset of MSRC dataset with totally 70 images. There are seven categories in this subset and each categories includes 10 images.

input-size	strategy a)			strategy b)			strategy c)			strategy d)		
	Car	Airplane	Horse	Car	Airplane	Horse	Car	Airplane	Horse	Car	Airplane	Horse
$k = 2$	85.8	77.2	74.1	84.4	78.3	73.9	85.9	77.3	74.0	85.7	77.3	74.1
$k = 3$	86.6	78.1	75.0	84.8	78.8	74.1	86.6	78.1	75.0	86.8	78.5	75.1
$k = 4$	-	-	-	85.6	78.9	75.1	86.9	78.6	75.4	87.0	78.8	75.5
$k = 5$	-	-	-	86.5	79.3	75.4	86.9	78.6	75.5	87.0	78.8	75.5
$k = 6$	-	-	-	86.4	79.0	75.4	86.9	78.7	75.4	86.9	78.7	75.5
$k = 7$	-	-	-	86.3	78.6	75.2	86.7	78.7	75.4	86.8	78.7	75.4
$k = 8$	-	-	-	86.1	78.9	75.3	86.7	78.7	75.5	86.8	78.7	75.5

Table 3.6: The results(Jaccard) of group segmentation with different sampling strategies and different numbers of input images on the Internet dataset. Our network can handle more than two images at a time, which generate better results than our network with paired input images.

Internet. Following [23, 7], we adopt the same subset of the Internet dataset with totally 300 images. There are three categories in this subset: horse, car, and airplane. Each category has 100 images.

iCoseg. There are 38 categories in the iCoseg dataset with total 643 images. A subset that contains 8 classes is commonly used to evaluate the generalizability of object co-segmentation methods.

We evaluate our proposed co-segmentation model and previous methods using two common evaluation metrics: *Precision* ($\mathcal{P}$) and *Jaccard Index* ($\mathcal{J}$.) The performance is reported under both metrics on these four co-segmentation benchmark datasets in our experiments. All the analysis experiments are performed on the Pascal VOC 2012 dataset.

3.3.3 Ablation Study

In this part, we investigate the effectiveness of each component in our proposed CycleSegNet by ablative analysis. We first create a baseline model which is a foreground prediction network where two branches do not exchange any information. Then we incorporate another two baseline methods M_{cat} and M_{mul} that exchanges information in the form of global representation. M_{cat} is to upsample the global vector and fuse it with the feature maps by concatenation, and M_{mul} is by multiplication. Then we add each of our proposed modules to the baseline model one by one. All the models are tested with the single-level feature encoder. The results are shown in Table. 3.1. As we can see, all the proposed modules are very effective in improving the performance over the baseline networks. Our cross region module (M_{cross}) significantly outperforms both baseline methods M_{cat} and M_{mul} by 9.9% and 9.6% in terms of Precision and 7.9% and 6.6% in terms of Jaccard, which shows that our design which exchanges information between local regions is more effective than previous methods using global representations. The cycle refinement module further boosts the Precision and Jaccard score by 4.7% and 2.2%, respectively.

Number of the refinement step. We report the performance of our network under different refinement steps N with fixed region size(2x2) in Table. 3.2. We also compare the average time to inference one image. We can see from the table, as the number of the refinement steps enlarges, the performance in terms of both Precision and Jaccard increases consistently while the increment in inference time is marginal. Particularly, after 7 steps of refinement, the network can improve the initial predictions by 3.6% and

3.5% in terms of Precision and Jaccard, respectively. In Figure. 3.4, we provide some visualization examples in the refinement process.

The size of the regional representations. We use the ROI pooling to get the regional representations of images in the cross region module. We next investigate how the region size influence the performance in Table. 3.3. As we can see, the best performance is observed when ROI pooling outputs a 2×2 representations. Directly using the original representations without ROI pooling does not yield better performance than regional representations. A possible reason is that the feature points in the original feature representations have a small effective field-of-view and thus lack the expressive power of abstract and semantic concepts. Using the ROI pooling can enlarge the field-of-view and provide context information in local regions, which is helpful in computing the correlation scores between regions for information exchanging.

Multi-level features. We compare the performance of our network using features from different layers in the encoder network in Table. 3.5. As can be observed, multi-level features can consistently boost the performance of our network with both the VGG16 backbone and the ResNet34 backbone. Moreover, using the features from the last three layers is better than using the features from the last two layers. This indicates that middle-level features are also helpful and can provide important information in the co-segmentation task.

Lovász-Softmax loss function. Additionally, we also conduct experiment on Lovász-Softmax loss function, which is commonly used in segmetation tasks. As is shown in Table. 3.5, compared with standard cross entropy loss, using the Lovász-Softmax loss function can brings about 1.7% and 1.0% improvement in terms of Precision and Jaccard, respectively.

Group segmentation We present a more detailed analysis about group segmentation in this part. The group segmentation intends to predict the common objects in a group of images. We apply different CNN based solutions in previous works to our network and then compare their performance.

Since the networks in most previous works for co-segmentation can only handle paired inputs, they often adopt various sampling strategies and fusion methods to undertake the group segmentation tasks, which can be applied to our network as well. An advantage of our network is that our network can handle more than two images at a time, as is described earlier, so the number of input images is flexible. However, directly sending all images in the group to our network is unrealistic due to the limited GPU memory. We therefore sample a small group of k images at a time and send them to our network for predictions. Then, we fuse all the predictions of an image as the final mask. Suppose that there are N images to be segmented in a group segmentation task and our network handles a tuple of k images at a time. We compare the following strategies to undertake group segmentation:

a) Following [3], we sample all possible k -element tuples to make predictions and average all the corresponding confidence maps of an image as the final prediction. This method can make use of all images in the group to make predictions for each image.

b) Following [77], we randomly divide all images into (N/k) small groups and each group has k images. In this case, each image is only tested once and the whole evaluation process is quick.

c) Following [33], to predict the mask of an image, we randomly sample 5 groups of images from the other $N - 1$ images, where each group has $k - 1$ images. Then, the target image joins each of these groups to make a prediction with our network, and we average

5 predictions to generate the final mask. The process is repeated for all N images.

d) Following [60], to predict the mask of an image, we uniformly split the other $N - 1$ images into T groups, where $T = (N - 1)/(k - 1)$ and each group has $k - 1$ images. Then, the target image joins each of these groups to make a prediction with our network, and we average T predictions to generate the final mask. The process is repeated for all N images.

Besides these solutions, there are also some previous works using graphs [41, 18] or recurrent models [60] to handle group segmentation, which, however, can not be applied to our network directly. Therefore, we compare the four solutions above with our network. We conduct experiments on the subset of Internet dataset, where each class has 100 images. For the solution **a**, generating all possible tuples for testing is unrealistic when k is large. Therefore, we only test the cases of $k = 2$ and $k = 3$. The result is shown in Table 3.6. As we can see, predicting with more than two input images is consistently better than predicting with paired input images with all four strategies, which shows the superiority of our network design for group segmentation. The best tuple size is 4 or 5, while further increasing the tuple size no longer boosts the performance, as is observed.

3.3.4 Comparison with the State-of-the-arts

Finally, we compare our proposed network with the state-of-the-art (SOTA) method on four popular benchmark datasets: Pascal VOC, MSRC, Internet and iCoseg dataset. We report the model performance for group segmentation as well as the pairwise co-segmentation. The results are shown in Table 3.7. As can be seen, our method outperforms all previous methods with a significant margin. Firstly, based on the training setting in [33, 7], we align the test split with these two studies for a fair comparison. For co-segmentation experiments on other datasets, we sample all possible pairs in each class for testing, as is done in previous works. On the challenging dataset Pascal VOC2012, the performance of our method outperforms the SOTA result by **10.9%** in terms of Jaccard (Table 3.7(a)); On the MSRC dataset, the performance of group segmentation and pair-wise segmentation outperform the SOTA by **4.3%** and **10.3%** respectively in terms of Jaccard (Table 3.7(c)); On the Internet dataset, the performance of group segmentation and pair-wise segmentation outperform the SOTA by **7.8%** and **5.9%** respectively in terms of Jaccard (Table 3.7(d)); Finally, on the challenging iCoseg dataset, where all testing classes are unseen in the training process, the performance of co-segmentation outperforms the SOTA by **4.8%**. Secondly, based on the training setting in recent methods [31, 77], our model also gets **2.6%**, **7.7%**, **2.2%**, **3.1%** gain in terms of Jaccard on Pascal VOC2010, MSRC, Internet and iCoseg dataset as shown in Table 3.7(b, c, d, f). Figure. 3.5 shows some qualitative results on these datasets.

3.4 Conclusion

In this work, we propose a novel and effective end-to-end approach for the task of object co-segmentation. The proposed Region Correspondence Module directly exchange information between local regions from two images, which shows obvious advantages over the baseline method transferring a global image representation. The Cycle Refinement Module that employs ConvLSTMs to progressively exchange information between images and update image representations can consistently improve the network predictions. Our algorithm can handle the inputs with paired images as well as grouped images with

the same network and parameters. The multi-level feature encoder can further boost the network performance effectively. Experiment results on four object co-segmentation benchmark datasets have proven the superiority of our design comparing with all existing approaches.

a) Pair-wise image co-segmentation

b) Group-wise image co-segmentation

Figure 3.1: Our network for image co-segmentation with paired or group input images. The main components in our network include a Cycle Refinement Module (CRM) for feature updating and a Region Correspondence Module(RCM) for information exchanging.

Figure 3.2: Architecture of Region Correspondence Module. $\oplus$ indicates element-wise add and $\otimes$ indicates matrix multiplication.

Figure 3.3: Our model variant utilizing multi-level features in the encoder for image co-segmentation. We apply the **CRM** module (described in Section. 3.2.2 and illustrated in Figure. 3.1) to features from different layers independently and fuse the resulting features using the Channel Attention module (CAM) proposed in [66] during the upsampling stage.

Figure 3.4: Predictions from different refinement steps. We decode the representations in each time step and the result shows that our proposed cycle refinement module can consistently improve the prediction.

Figure 3.5: Qualitative results(Jaccard) of our network on the PASCAL VOC, MSRC, Internet, iCoseg dataset.

VOC2012(pair)	$\mathcal{P}(\%)$	$\mathcal{J}(\%)$
CA[7]	-	59.2
FCA[7]	-	59.4
CSA[7]	-	59.8
DOCS[33]	94.2	64.5
CycleSegNet(Ours)	95.8	75.4

(a) Pairwise co-segmentation results on the **Pascal VOC2012** dataset

VOC2010(group)	$\mathcal{P}(\%)$	$\mathcal{J}(\%)$
Quan <i>et al.</i> [41]	89.0	52.0
Jerripothula <i>et al.</i> [25]	85.2	45.0
Jerripothula <i>et al.</i> [27]	80.1	40.0
Li <i>et al.</i> [31]	94.1	63.0
Zhang <i>et al.</i> [77]	94.9	71.0
CycleSegNet(Ours)	96.8	73.6

(b) group-wise co-segmentation results on the **Pascal VOC2010** dataset

MSRC	train setting	group		pair	
		$\mathcal{P}(\%)$	$\mathcal{J}(\%)$	$\mathcal{P}(\%)$	$\mathcal{J}(\%)$
Vicente <i>et al.</i> [56]	-	90.2	70.6	-	-
Wang <i>et al.</i> [58]	-	92.2	-	-	-
Rubinstein <i>et al.</i> [47]	-	92.2	74.7	-	-
Faktor <i>et al.</i> [14]	-	92.0	77.0	-	-
Chen <i>et al.</i> [7]	VOC2012	-	73.9	96.6	76.5
Li <i>et al.</i> [33]	VOC2012	95.4	82.9	-	-
CycleSegNet(Ours)	VOC2012	97.9	87.2	97.3	86.8
Zhang <i>et al.</i> [77]	COCO	95.2	81.9	-	-
Li <i>et al.</i> [31]	COCO	-	-	-	-
CycleSegNet(Ours)	COCO	97.6	89.6	97.2	88.2

(c) Results on the **MSRC** dataset.

Internet	train setting	group				pair			
		Airplane	Car	Horse	Avg. $\mathcal{J}(\%)$	Airplane	Car	Horse	Avg. $\mathcal{J}(\%)$
Rubinstein <i>et al.</i> [47]	-	55.8	64.4	51.6	57.3	-	-	-	-
Quan <i>et al.</i> [41]	-	56.3	66.8	58.1	60.4	-	-	-	-
Jerripothula <i>et al.</i> [25]	-	61.0	71.0	60.0	64.0	-	-	-	-
Chen <i>et al.</i> [10]	-	65.0	82.0	63.0	70.0	-	-	-	-
Yuan <i>et al.</i> [68]	VOC2012	66.0	72.0	65.0	67.7	-	-	-	-
Li <i>et al.</i> [33]	VOC2012	65.4	82.8	69.4	72.6	-	-	-	-
Chen <i>et al.</i> [7]	VOC2012	-	-	-	-	71.4	79.9	68.0	73.1
CycleSegNet(Ours)	VOC2012	78.8	87.0	75.5	80.4	77.2	85.8	74.1	79.0
Zhang <i>et al.</i> [77]	COCO	69.6	82.5	70.2	74.1	-	-	-	-
Li <i>et al.</i> [31]	COCO	83.0	93.0	76.0	84.0	-	-	-	-
CycleSegNet(Ours)	COCO	84.1	91.6	82.9	86.2	83.2	88.5	80.3	84.0

(d) Results on **Internet** dataset

iCoseg(pair)	train setting	bear2	brownbear	cheetah	elephant	helicopter	hotballoon	panda1	panda2	Avg. $\mathcal{J}(\%)$
Rubinstein <i>et al.</i> [47]	-	65.3	73.6	69.7	68.8	80.3	65.7	75.9	62.5	70.2
Jerripothula <i>et al.</i> [26]	-	70.1	66.2	75.4	73.5	76.6	76.3	80.6	71.8	73.8
Faktor <i>et al.</i> [14]	-	72.0	66.2	75.4	73.5	76.6	76.3	80.6	71.8	73.8
Jerripothula <i>et al.</i> [25]	-	67.5	72.5	78.0	79.9	80.0	80.2	72.2	61.4	70.4
Li <i>et al.</i> [33]	VOC2012	88.3	92.0	68.8	84.6	79.0	91.7	82.6	86.7	84.2
Chen <i>et al.</i> [7]	VOC2012	88.3	91.5	71.3	84.4	76.5	94.0	91.8	90.3	86.0
CycleSegNet(Ours)	VOC2012	92.1	94.8	84.5	90.2	80.2	95.8	94.5	94.1	90.8

(e) Results on the **iCoseg** dataset

iCoseg(group)	train setting	bear2	brownbear	cheetah	elephant	helicopter	hotballoon	panda1	panda2	Avg. $\mathcal{J}(\%)$
Zhang <i>et al.</i> [77]	COCO	91.1	89.6	88.6	90.0	76.4	94.2	90.4	87.5	89.2
CycleSegNet(Ours)	COCO	92.8	94.1	89.1	91.6	85.2	95.1	94.8	94.1	92.1

(f) Results on the **iCoseg** dataset

Table 3.7: Comparison with the state-of-the-art performance of pairwise co-segmentation and group segmentation tasks on four benchmark datasets: **Pascal VOC**, **MSRC**, **Internet** and **iCoseg**. Our model achieves new state-of-the-art performance on all datasets.

Chapter 4

Weakly-supervised One-shot Segmentation

4.1 Introduction

Image segmentation is one of the fundamental tasks in computer vision. Deep learning methods have greatly accelerated its development in recent years. By fitting on large-scale datasets containing thousands of annotated images, such as PASCAL VOC [12] and MS COCO [35], the models gain the ability to segment objects. However, pixel-level labels can be very expensive in some situations, and the current approaches lack the ability to segment objects from new categories after training on pre-defined classes without any extra finetuning. On the contrary, humans can easily segment new categories of objects without looking too much at the samples, which means the great potential improvement of deep neural networks.

Few-shot image segmentation, which aims to predict unseen categories with a few annotated images after training, is essential for realistic applications. Compared with the common segmentation approaches, which predicts masks as results for given input images, few-shot segmentation focuses on mining the correlation between the query and the support image, which greatly improves the generalization ability given the limited data and annotations. Researchers have already investigated this topic in [50, 70, 72, 59, 81].

However, existing methods in few-shot object segmentation have mainly relied on manually labeled segmentation masks. Limited research has been conducted on using weak supervision in the training process of few-shot segmentation. This is the key setting (one-shot and bounding boxes level supervision) this thesis attempts to explore. The main difficulties of this task are limited data and weak supervision, assuming pixel-wise annotations are not provided in the training process. The set-up is very practical and common in the real world, since the ability to segment objects from one or a few reference images is a natural case for humans.

In recent time, many researches are devoted to weakly supervised semantic segmentation, which only require weak supervisions, such as scribbles, bounding boxes, or even image-level classification labels, aiming to reduce the performance gap between full and weak supervisions. Most of the advanced weakly supervised semantic segmentation networks exploit the class activation map(CAM) [83], which has been proven to discover and extract the local discriminative parts of the target object. The discriminative regions highlighted by CAMs will be used as seed and expanded to the whole object area.

Since in the one-shot or few shot cases, the testing classes and the training classes are non-overlapping, thus these methods are not applicable directly. However, we can utilize these weakly-supervised methods in the bounding box regions to generate pseudo masks for our one-shot training and provide more accurate object information.

To mitigate the labeling requirements, we present our method for weakly-supervised one-shot segmentation. The major difference between most of previous one-shot methods and our weakly-supervised solution is that we only need bounding box level supervision in the one-shot training process instead of high-cost pixel-wise annotations. Since the bounding box annotations can not provide accurate object information as the pixel-wise mask labels, we apply weakly-supervised semantic segmentation approaches to obtain pseudo masks in the bounding box regions. Then we take these refined labels to supervise the training process of our one-shot segmentation network. In order to effectively obtain the correlation information between support image and query image, we take the Graph Attention Unit from [72] to model query feature map and support feature map in different levels. In addition, we propose Design Extraction Module to fully extract the detailed information, and Feature Fusion Module to merge the feature correlation with the detailed information for network prediction.

To validate our network designs, we implement various experiments on the dataset PASCAL-5ⁱ. Our weakly-supervised method delivers comparable performance with the state-of-the-art fully-supervised one-shot approaches. The major contributions of our work are summarized as follows:

- We propose a novel low-cost attempt for one-shot segmentation methods relying on pixel-level supervision to support weak (bounding box label) supervision.
- Based on the bounding box annotations, we reference the recent weakly supervised segmentation approach to generate pseudo mask labels to supervise the training process of one-shot segmentation.
- We propose a Detail Extraction Module and Feature Fusion Module in the one-shot segmentation model to provide multi-level detailed information and integrate the feature correlation with the detailed information.
- We extensively evaluate our proposed weakly-supervised method on the PASCAL-5ⁱ dataset and narrow down the performance gap between weakly-supervised and full-supervised one-shot segmentation methods.

4.2 Methodology

In this section, we present our design for the task of weakly-supervised one-shot segmentation. We begin with the problem formulation of weakly-supervised one-shot segmentation task, then introduce the generation of one-shot training masks by weakly-supervised semantic segmentation method. We follow the recent weakly-supervised semantic segmentation network [62] to obtain pseudo object masks, which utilizes a self-supervised equivariant regularization and random walk to refine the class activation map(CAM). Finally, we describe our proposed weakly-supervised one-shot segmentation network.

Figure 4.1: The whole architecture of our one-shot segmentation network. DEM indicates Detail Extraction Module and FFM represents Feature fusion Module. Given a query image and a support image, we first employ a weight-shared CNN to extract their features in different levels. Then we utilize the graph-based method proposed in [72] to model the correspondence between the middle-level features from support image and query image, and our proposed Detail Extraction Module to capture object detail information based on low-level feature map. After that, the detailed information and correspondence information are processed by Feature Fusion Module and further upsampled to generate the predicted mask.

4.2.1 Problem Definition

The previous literature focused mainly on using strongly labeled pixel-level segmentation masks for the few examples in the support set. It is labour intensive and impractical to provide such annotations for every single novel class. In our set-up, we use weaker annotation (bounding box) to supervise the training of few-shot segmentation. Assume that our model is trained on the class set C_{train} , the objective of our task is that given one annotated image of new classes C_{test} , the trained model can discover the objects of new classes C_{test} in query image and make the prediction. We only use weak annotations in the training dataset instead of pixel-level labels. Intuitively, we train the model to have the ability that for a new class $c \notin C_{train}$, our model is capable to find out relevant areas in both support images and query images and output the predicted masks of this new class.

Our training and testing process is based on episodic paradigm [57], but utilizes weak supervision instead of pixel-wise annotations in the training process. To be specific, in this one-shot learning task, each episode consists of: 1) a support (training) set $S = \{(x_s^i, y_s^i(c))\}$, where $x_s^i \in R^{H_i \times W_i \times 3}$ is the support RGB image and $y_s^i(c) \in R^{H_i \times W_i}$ is the corresponding binary bounding-box level mask for class c ; and 2) a query (training) set $Q = \{x_q, y_q(c)\}$ where x_q is the query image and $y_q(c)$ is the binary bounding-box level mask for class c in the query image. Given the support set S and the query image x_q

as input, the model outputs the predicted mask $\hat{y}_q(c)$ for category c in the query image. We ensure that the training categories and testing categories are non-overlapping, and no repeated image-label pairs exist in both training set and testing set.

4.2.2 Pseudo Masks Generation

To generate pseudo mask labels based on bounding-boxes, we first train a weakly-supervised segmentation network to obtain the class activation map(CAM), and then refine the CAM to define the object regions as pseudo masks for one-shot segmentation following SEAM [62]. Based on the one-shot training samples and their corresponding bounding boxes, we first takes all their instances belonging to the training classes to construct a training dataset for weakly-supervised semantic segmentation. To reduce the loss of objects' edge information, we randomly expand the size of the box regions and then crop the expanded regions as training images.

SEAM [62] introduces equivariant regularization and pixel correlation module to improve the quality of segmentation masks. Fig. 4.2 shows the whole structure of SEAM network. It utilizes a shared-weight Siamese network to process two branches of inputs. A branch applies the spatial affine transformation to the input image while the other applies the same transformation to the output prediction. At the end of the network, the Pixel Correlation Module(PCM) is proposed to capture more context information and refine the activation map predictions, which can be regarded as a variant of self-attention mechanism. PCM computes a similarity matrix by inner-product between the embedding space of two pixels and then uses ReLU activation function to suppress the irrelevant values. The final CAM is obtained by the weighted sum of the original CAM with normalized similarities. From Fig. 4.2, we can find that after PCM, the refined CAM can be more accurate than the original CAM. The equivariant regularization is reflected in the loss functions, which contains classification loss, ER loss, and ECR loss. ER loss is proposed to guarantee the consistency of CAMs over various affine transformations. ECR loss is used to introduce equivariant regularization between the origin CAMs and revised CAMs.

During inference, we apply horizontal flip and multi-scale test to the input images, and then aggregate the results of the processed images to make the prediction, which is a common practice to enhance the equivariance of predicted masks and improve the segmentation performance. To further refine the pseudo masks, we also refer to the work of [1] to train an AffinityNet using our revised CAM. AffinityNet is proposed to generate semantic affinity matrix between a pair of image coordinates within a small area, which is combined with random walk to generate the pseudo mask labels.

4.2.3 One-shot Segmentation Network

Based on the recent network [72] which proposes a graph-based method to model the correspondence between query image and support image, we further introduce Detail Extraction Module(DEM) to extract the low-level detailed information, and Feature Fusion Module to integrate the detailed information and correspondence information to generate the predicted masks. The whole architecture of our proposed few-shot segmentation network is as shown in Fig.4.1.

Detail Extraction Module(DEM) The goal of our proposed detail extraction module is to extract different levels of details. We use DEM to process the low level

Figure 4.2: Architecture of SEAM [62]

feature map from the backbone in parallel with different receptive fields. As illustrated in Fig.4.3, given a input feature map $x \in R^{C \times H \times W}$, we use four group convolutions with different kernel sizes to extract multi-level detail information $\{x_1, x_2, x_3, x_4\} \in R^{C' / 4 \times H \times W}$ and then concatenate these feature maps as output. We enlarge the kernel size from k_1 to k_4 and also expand the number of group from g_1 to g_4 . Thus compared with the traditional convolution, different types of kernels can provide more reliable detailed information about larger or smaller objects and richer context information, also help to boost the segmentation performance of our network without increasing additional costs or model sizes. Also, the combination of kernel size and the number of groups in each convolution block is flexible and we can adjust it in different settings.

Feature Fusion Module(FFM) In the previous model [72], the middle-level features are used to undertake graph reasoning to establish the correspondence between query images and support images. And as mentioned above, we utilize Detail Extraction Module to complement the detailed information. Then these two features are processed by our proposed Feature Fusion Module(FFM) before generating the final segmentation masks. Our FFM is designed to adjust the weights of features from different stages to enhance the consistency as illustrated in Fig. 4.4. Given two feature maps $f_1 \in R^{C \times H \times W}$, $f_2 \in R^{C \times H/2 \times W/2}$, we first upsample f_2 to the same resolution with f_1 , then concatenate them to learn a channel-wise weights α to explore the significance of each channel. α is then applied to the origin f_1 to extract the discriminative features and suppress the less significant features. The output of FFM is the summation of the weighted f_1 and upsampled f_2 . Our FFM shares similar structure with the Channel Attention Block proposed in literature [66], which is used to enhance the consistency when combining the features of adjacent stages.

$$\alpha = \text{Sigmoid}(\text{Conv}_{1 \times 1}(\text{ReLU}(\text{Conv}_{1 \times 1}(\text{Pool}(\text{concat}(f_1, \text{Up}(f_2))))))), \quad (4.1)$$

$$\text{output} = \alpha * f_1 + \text{Up}(f_2), \quad (4.2)$$

Figure 4.3: Architecture of our proposed Detail Extraction Module.

Methods	mean IoU
Ours-origin	41.8
Ours-pseudo	51.5

Table 4.1: Ablation experiments on the Pseudo Masks Generation process. Ours-origin indicates using bounding box labels as supervision in one-shot training, and Ours-pseudo indicates utilizing generated pseudo labels to supervise one-shot training. Both of the methods use our proposed one-shot segmentation network.

4.3 Experiment Results

4.3.1 Implementation Details

Regarding the Pseudo Masks Generation, we fully reference the approach SEAM [62] to train the weakly-supervised segmentation network but use the training instances from one-shot segmentation datasets as training samples.

When training the weakly-supervised one-shot segmentation network, we adopt binary Cross-Entropy loss to guide and constrain our model. We use momentum SGD optimizer and train the model for 200 epochs with the learning rate of 0.0025 and batch size of 12. The momentum parameter is set to 0.9 and weight decay is 0.0005. We utilize random crop, random scale, and random horizontal flip as augmentations on the support

Methods	mean-IoU
Ours-init	50.0
Ours-Conv	50.3
Ours-DEM	51.5

Table 4.2: Ablation experiments on the Detail Extraction Module(DEM). Ours-init indicates discarding DEM in our proposed model, Ours-Conv means using a convolution layer with kernel size 3x3 to replace DEM, and Ours-DEM represents our proposed model.

Methods	mean IoU
Ours-sum	50.1
Ours-concat	50.5
Ours-FFM	51.5

Table 4.3: The results of different fusion operations on the PASCAL-5ⁱ dataset. Ours-sum, Ours-concat, Ours-FFM represent using element-wise summation, concatenation and our proposed FFM to fuse feature in different levels respectively.

Methods	Support	Query	mean IoU
Ours			51.5
Ours	✓		52.9
Ours		✓	51.8
Ours	✓	✓	53.1

Table 4.4: Ablation experiments on the Muti-Scale Evaluation. All predictions are rescaled to the same image size and merged by average.

Methods	supervision	fold0	fold1	fold2	fold3	mean
Reduced-DFCN8s[52]	pixel-wise	39.2	48.0	39.3	34.2	40.2
OSLSM[50]	pixel-wise	33.6	55.3	40.9	33.5	40.8
CoFCN[42]	pixel-wise	36.7	50.6	44.9	32.4	41.1
GN[43]	pixel-wise	37.5	50.0	44.1	39.4	41.4
AMP[53]	pixel-wise	41.9	50.2	46.7	34.7	43.4
SG-One[81]	pixel-wise	40.2	58.4	48.4	38.4	46.3
PANet[59]	pixel-wise	42.3	58.0	51.1	41.2	48.1
CANet[70]	pixel-wise	52.5	65.9	51.3	51.9	55.4
PGNet[72]	pixel-wise	56.0	66.9	50.6	50.4	56.0
Ours	bounding box	47.0	62.6	48.3	47.9	51.5
Ours-MS	bounding box	48.7	64.1	49.9	49.5	53.1

Table 4.5: mean IoU of one-shot segmentation on the PASCAL-5ⁱ dataset under pixel-wise supervision and bounding box supervision. Ours-MS represents our proposed method with Multi-Scale input during testing

Figure 4.4: Architecture of our proposed Feature Fusion Module(FFM).

images during training. In our proposed Detail Extraction Module, the specific kernel size k_1, k_2, k_3, k_4 we use is 3, 5, 7, 9, and the number of group g_1, g_2, g_3, g_4 is 1, 4, 8, 16.

4.3.2 Datasets and Evaluation Metrics.

The dataset used for the experiments is PASCAL-5ⁱ proposed in [50], which is built on the whole PASCAL VOC 2012 dataset and additional annotation files from the SBD [19] dataset. In our experiment, We adopt the same split and the same random seed proposed by [50] to guarantee fair comparison. PASCAL-5ⁱ is splitted into 4 groups and each group contains 5 categories. For each split, we choose one group as testing set, the other three groups are used as training set. In the test stage, we use the same random seed and randomly sample 1,000 image pairs in each test fold. Different from the previous one-shot segmentation approach, we use weak supervision(bounding box level annotations) instead of pixel-wise annotations in training.

We choose mean IoU as the evaluation metric. The Intersection over Union (IoU) metric calculates the percentage of overlap areas between mask labels and predicted masks, which is very common to evaluate the performance of segmentation approaches. The mean IoU is the average IoU of all testing categories. To be specific, in the test stage, we first calculate the IoU score for each testing category, and then average the IoU score of five categories as mean IoU for each split.

4.3.3 Ablation Study

In this part, we investigate the effectiveness of each component in our proposed one-shot segmentation network by ablation experiments on the PASCAL-5ⁱ dataset. There are four parts for ablation studies: Pseudo Masks Generation, Detail Extraction Module, Feature Fusion Module and Multi-Scale Evaluation. The ablation experiments help us to confirm which module significantly contributes and determine whether these modules are necessary. We use average mean IoU over 4 splits on the PASCAL-5ⁱ dataset to evaluate the performance.

Figure 4.5: Qualitative results of our Pseudo Masks Generation. The first, third, and fifth columns are the input images. The second, fourth, and sixth columns are our generated pseudo masks.

Figure 4.6: Qualitative results of our weakly-supervised one-shot segmentation network. The first, third, and fifth columns are the input images. The second, fourth, and sixth columns are our predicted masks.

We first evaluate the significance of our **Pseudo Masks Generation** process. We conduct experiments on two setups: use original bounding box labels(origin) and pseudo mask labels(pseudo). As shown in Table 4.1, the results indicate that the pseudo masks contribute 9.7% improvement of mean IoU compared with original bounding box masks. We believe that the generated pseudo masks is an essential step in our approach, providing accurate information about the target object areas. Some of our generated pseudo masks shown in Figure 4.5 are close to the pixel-wise segmentation labels.

To illustrate the ability of **Detail Extraction Module**, we conduct the experiment with three setups: without Detail Extraction Module(init), use one convolution layer with kernel size 3x3 to replace Detail Extraction Module(Conv), and use our Detail Extraction Module(DEM). As is shown in Table 4.2, our DEM contributes 1.5% improvement of mean IoU compared without this module and 1.2% improvement compared with simple convolution layer. The experiment results indicate that Detail Extraction Module can capture more complementary information in the one-shot segmentation task.

We compare different fusion approaches to prove that our **Feature Fusion Module** is better when fusing the features from different stages. We conduct the experiment with three setups: summation, concatenation, and our proposed Feature Fusion Module. In Table 4.3, the results indicate that our FFM has better performance to fuse the features, and can achieve more significant improvement in the task of one-shot segmentation.

Additionally, We also conduct experiment on **Multi-Scale Evaluation**, which is commonly used in segmentation literature. To be specific, the input support image and query image are rescaled by [0.7, 1.0, 1.3]. Then we rescale the predicted masks back to the original size by bilinear interpolation and average them as the final result. As is shown in Table 4.4, multi-scale evaluation brings about 1.6% mean IoU improvement.

4.3.4 Comparison to the Pixel-wise Annotation

We compare our methods with the existing methods in pixel-wise level annotation and bounding-box level annotation on the dataset PASCAL-5ⁱ in Table 4.5. Without the utilization of pixel-wise level segmentation mask, our method with the bounding box level supervision performs(53.1%) close to the current state-of-the-art fully-supervised methods(56.0%) in the one-shot case. Simultaneously, our weakly-supervised one-shot segmentation method even outperforms a number of fully-supervised one-shot segmentation methods. Figure 4.6 shows some results of our weakly-supervised one-shot segmentation network. We can find that our network has the ability to make accurate predictions of the query images with only one labeled example image.

4.4 Conclusion

In this thesis, we propose a novel and effective method for the task of one-shot segmentation in weak supervision. Compared with previous fully supervised one-shot segmentation methods, we only require bounding box level annotations during training instead of expensive pixel-wise labels. We first adopt the recent weakly-supervised semantic segmentation approach to generate the pseudo mask labels. Then in our one-shot segmentation network, we utilize Detail Extraction Module to extract the detailed information with different receptive fields without increasing additional costs, and Graph Attention Unit to model the correlation between support image and query image. Finally, our Feature

Fusion Module is designed to integrate the detailed information and correspondence information for network prediction. Experiment results on dataset PASCAL-5ⁱ show the effectiveness of our approach, which lessen the performance gap between bounding box level supervision and pixel-wise annotations.

Chapter 5

Conclusion and Future Work

5.1 Conclusion

In this thesis, we focus on tackling two problems: object co-segmentation and weakly-supervised one-shot segmentation. Object co-segmentation aims to utilize joint information from multiple inputs and segment the similar objects. Weakly-supervised one-shot segmentation aims to utilize weak supervision (bounding box level supervision in this thesis) in the one-shot domain.

As for object co-segmentation, we propose a novel and effective network named CycleSegNet. In order to exchange information between local regions from two images, we design a Region Correspondence Module that shows obvious advantages over the baseline method transferring a global image representation. Our proposed Cycle Refinement Module that employs ConvLSTMs to progressively exchange information between images and update image representations can consistently improve the network predictions. Furthermore, our algorithm can handle the inputs with paired images as well as grouped images with the same network and parameters. The multi-level feature encoder can further boost the network performance effectively. Experiment results on four object co-segmentation benchmark datasets validate our design and our method surpass the existing state-of-the-art co-segmentation models.

As for weakly-supervised one-shot segmentation, we alleviate the requirement of pixel-wise annotations and only need bounding box level supervision in the training process. Compared with previous fully supervised one-shot segmentation methods, we adopt the recent weakly-supervised semantic segmentation approach to generate the pseudo mask labels as supervision. Then in our one-shot segmentation network, we utilize Detail Extraction Module to extract the detailed information with different receptive fields without increasing additional costs, and Graph Attention Unit to model the correlation between support image and query image. Finally, our Feature Fusion Module is designed to integrate the detailed information and correspondence information for network prediction. Experiment results on dataset PASCAL-5ⁱ show the effectiveness of our approach, which lessen the performance gap between bounding box level supervision and pixel-wise annotations.

5.2 Future Work

Despite the active studies in the field of object co-segmentation and one-shot segmentation, there are still some limitations with the proposed method, and further studies and improvements are demanded.

As for the object co-segmentation task, collecting a large number of manually labeled datasets is time-consuming and expensive. Thus weakly-supervised or unsupervised approaches for object co-segmentation will be our future work.

As for weakly-supervised one-shot segmentation, there is still much room to improve for weakly-supervised approaches. In our following work, we will attempt to further improve the segmentation performance under bounding box level supervision and even image-level supervision. In addition, how to make full use of the bounding box information in the weakly-supervised one shot domain is worth exploring.

Bibliography

- [1] Jiwoon Ahn and Suha Kwak. “Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 4981–4990.
- [2] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. “Segnet: A deep convolutional encoder-decoder architecture for image segmentation”. In: *IEEE transactions on pattern analysis and machine intelligence* 39.12 (2017), pp. 2481–2495.
- [3] Sayan Banerjee et al. “CoSegNet: Image co-segmentation using a conditional siamese convolutional network”. In: *Proc. 28th Int. Joint Conf. Artif. Intell.* 2019, pp. 673–679.
- [4] Dhruv Batra et al. “icoseg: Interactive co-segmentation with intelligent scribble guidance”. In: *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE. 2010, pp. 3169–3176.
- [5] Maxim Berman, Amal Rannen Triki, and Matthew B Blaschko. “The lovász-softmax loss: a tractable surrogate for the optimization of the intersection-over-union measure in neural networks”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 4413–4421.
- [6] Kai-Yueh Chang, Tyng-Luh Liu, and Shang-Hong Lai. “From co-saliency to co-segmentation: An efficient and fully unsupervised energy minimization model”. In: *CVPR 2011*. IEEE. 2011, pp. 2129–2136.
- [7] Hong Chen, Yifei Huang, and Hideki Nakayama. “Semantic aware attention based deep object co-segmentation”. In: *Asian Conference on Computer Vision*. Springer. 2018, pp. 435–450.
- [8] Hsin-Yi Chen, Yen-Yu Lin, and Bing-Yu Chen. “Co-segmentation guided hough transform for robust feature matching”. In: *IEEE transactions on pattern analysis and machine intelligence* 37.12 (2015), pp. 2388–2401.
- [9] Liang-Chieh Chen et al. “Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs”. In: *IEEE transactions on pattern analysis and machine intelligence* 40.4 (2017), pp. 834–848.
- [10] Yun-Chun Chen et al. “Show, Match and Segment: Joint Learning of Semantic Matching and Object Co-segmentation”. In: *arXiv preprint arXiv:1906.05857* (2019).
- [11] Jia Deng et al. “Imagenet: A large-scale hierarchical image database”. In: *2009 IEEE conference on computer vision and pattern recognition*. Ieee. 2009, pp. 248–255.

- [12] Mark Everingham et al. “The pascal visual object classes (voc) challenge”. In: *International journal of computer vision* 88.2 (2010), pp. 303–338.
- [13] Mark Everingham et al. “The pascal visual object classes challenge: A retrospective”. In: *International journal of computer vision* 111.1 (2015), pp. 98–136.
- [14] Alon Faktor and Michal Irani. “Co-segmentation by composition”. In: *Proceedings of the IEEE International Conference on Computer Vision*. 2013, pp. 1297–1304.
- [15] Deng-Ping Fan et al. “Re-thinking co-salient object detection”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2021).
- [16] Jun Fu et al. “Stacked deconvolutional network for semantic segmentation”. In: *IEEE Transactions on Image Processing* (2019).
- [17] Junwei Han et al. “A unified metric learning-based framework for co-saliency detection”. In: *IEEE Transactions on Circuits and Systems for Video Technology* 28.10 (2017), pp. 2473–2483.
- [18] Junwei Han et al. “Robust object co-segmentation using background prior”. In: *IEEE Transactions on Image Processing* 27.4 (2017), pp. 1639–1651.
- [19] Bharath Hariharan et al. “Simultaneous detection and segmentation”. In: *European Conference on Computer Vision*. Springer. 2014, pp. 297–312.
- [20] Kaiming He et al. “Deep residual learning for image recognition”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.
- [21] Kaiming He et al. “Mask r-cnn”. In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 2961–2969.
- [22] Sepp Hochreiter and Jürgen Schmidhuber. “Long short-term memory”. In: *Neural computation* 9.8 (1997), pp. 1735–1780.
- [23] Kuang-Jui Hsu, Yen-Yu Lin, and Yung-Yu Chuang. “Co-attention CNNs for Unsupervised Object Co-segmentation.” In: *IJCAI*. 2018, pp. 748–756.
- [24] Tao Hu et al. “Attention-based multi-context guiding for few-shot semantic segmentation”. In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 33. 01. 2019, pp. 8441–8448.
- [25] Koteswar Rao Jerripothula, Jianfei Cai, and Junsong Yuan. “Image co-segmentation via saliency co-fusion”. In: *IEEE Transactions on Multimedia* 18.9 (2016), pp. 1896–1909.
- [26] Koteswar Rao Jerripothula et al. “Automatic image co-segmentation using geometric mean saliency”. In: *2014 IEEE International Conference on Image Processing (ICIP)*. IEEE. 2014, pp. 3277–3281.
- [27] Koteswar Rao Jerripothula et al. “Object co-skeletonization with co-segmentation”. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE. 2017, pp. 3881–3889.
- [28] Longlong Jing, Yucheng Chen, and Yingli Tian. “Coarse-to-fine semantic segmentation from image-level labels”. In: *IEEE Transactions on Image Processing* 29 (2019), pp. 225–236.
- [29] Armand Joulin, Francis Bach, and Jean Ponce. “Discriminative clustering for image co-segmentation”. In: *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE. 2010, pp. 1943–1950.

- [30] Diederik P Kingma and Jimmy Ba. “Adam: A method for stochastic optimization”. In: *arXiv preprint arXiv:1412.6980* (2014).
- [31] Bo Li et al. “Group-Wise Deep Object Co-Segmentation With Co-Attention Recurrent Neural Network”. In: *Proceedings of the IEEE International Conference on Computer Vision*. 2019, pp. 8519–8528.
- [32] Chongyi Li et al. “ASIF-Net: Attention steered interweave fusion network for RGB-D salient object detection”. In: *IEEE transactions on cybernetics* 51.1 (2020), pp. 88–100.
- [33] Weihao Li, Omid Hosseini Jafari, and Carsten Rother. “Deep object co-segmentation”. In: *Asian Conference on Computer Vision*. Springer. 2018, pp. 638–653.
- [34] Di Lin et al. “Scribblesup: Scribble-supervised convolutional networks for semantic segmentation”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016, pp. 3159–3167.
- [35] Tsung-Yi Lin et al. “Microsoft coco: Common objects in context”. In: *European conference on computer vision*. Springer. 2014, pp. 740–755.
- [36] Weide Liu et al. “CRNet: Cross-Reference Networks for Few-Shot Segmentation”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2020.
- [37] Jonathan Long, Evan Shelhamer, and Trevor Darrell. “Fully convolutional networks for semantic segmentation”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 3431–3440.
- [38] Lane McIntosh et al. “Recurrent segmentation for variable computational budgets”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 2018, pp. 1648–1657.
- [39] Lopamudra Mukherjee, Vikas Singh, and Jiming Peng. “Scale invariant cosegmentation for image groups”. In: *CVPR 2011*. IEEE. 2011, pp. 1881–1888.
- [40] Armin Mustafa and Adrian Hilton. “Semantically coherent co-segmentation and reconstruction of dynamic scenes”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017, pp. 422–431.
- [41] Rong Quan et al. “Object co-segmentation via graph optimized-flexible manifold ranking”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016, pp. 687–695.
- [42] Kate Rakelly et al. “Conditional networks for few-shot semantic segmentation”. In: (2018).
- [43] Kate Rakelly et al. “Few-shot segmentation propagation with guided networks”. In: *arXiv preprint arXiv:1806.07373* (2018).
- [44] Shaoqing Ren et al. “Faster r-cnn: Towards real-time object detection with region proposal networks”. In: *Advances in neural information processing systems*. 2015, pp. 91–99.
- [45] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. “U-net: Convolutional networks for biomedical image segmentation”. In: *International Conference on Medical image computing and computer-assisted intervention*. Springer. 2015, pp. 234–241.

- [46] Carsten Rother et al. “Cosegmentation of image pairs by histogram matching-incorporating a global constraint into mrfs”. In: *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’06)*. Vol. 1. IEEE. 2006, pp. 993–1000.
- [47] Michael Rubinstein et al. “Unsupervised joint object discovery and segmentation in internet images”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2013, pp. 1939–1946.
- [48] Jose C Rubio et al. “Unsupervised co-segmentation through region matching”. In: *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE. 2012, pp. 749–756.
- [49] Tomas Sakinis et al. “Interactive segmentation of medical images through fully convolutional neural networks”. In: *arXiv preprint arXiv:1903.08205* (2019).
- [50] Amirreza Shaban et al. “One-Shot Learning for Semantic Segmentation”. In: 2017.
- [51] Jamie Shotton et al. “Textonboost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation”. In: *European conference on computer vision*. Springer. 2006, pp. 1–15.
- [52] Mennatullah Siam and Boris Oreshkin. “Adaptive masked weight imprinting for few-shot segmentation”. In: (2019).
- [53] Mennatullah Siam, Boris N Oreshkin, and Martin Jagersand. “AMP: Adaptive masked proxies for few-shot segmentation”. In: *Proceedings of the IEEE International Conference on Computer Vision*. 2019, pp. 5249–5258.
- [54] Karen Simonyan and Andrew Zisserman. “Very deep convolutional networks for large-scale image recognition”. In: *arXiv preprint arXiv:1409.1556* (2014).
- [55] Guolei Sun et al. “Mining cross-image semantics for weakly supervised semantic segmentation”. In: *European Conference on Computer Vision*. Springer. 2020, pp. 347–365.
- [56] Sara Vicente, Carsten Rother, and Vladimir Kolmogorov. “Object cosegmentation”. In: *CVPR 2011*. IEEE. 2011, pp. 2217–2224.
- [57] Oriol Vinyals et al. “Matching networks for one shot learning”. In: *Advances in neural information processing systems*. 2016, pp. 3630–3638.
- [58] Fan Wang, Qixing Huang, and Leonidas J Guibas. “Image co-segmentation via consistent functional maps”. In: *Proceedings of the IEEE International Conference on Computer Vision*. 2013, pp. 849–856.
- [59] Kaixin Wang et al. “Panet: Few-shot image semantic segmentation with prototype alignment”. In: *Proceedings of the IEEE International Conference on Computer Vision*. 2019, pp. 9197–9206.
- [60] Wenguan Wang et al. “Zero-shot video object segmentation via attentive graph neural networks”. In: *Proceedings of the IEEE International Conference on Computer Vision*. 2019, pp. 9236–9245.
- [61] Xiang Wang et al. “Weakly-supervised semantic segmentation by iteratively mining common object features”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 1354–1362.

- [62] Yude Wang et al. “Self-supervised Equivariant Attention Mechanism for Weakly Supervised Semantic Segmentation”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, pp. 12275–12284.
- [63] Lina Wei et al. “Group-wise Deep Co-saliency Detection”. In: *IJCAI*. 2017.
- [64] Yunchao Wei et al. “Object region mining with adversarial erasing: A simple classification to semantic segmentation approach”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 1568–1576.
- [65] Sanghyun Woo et al. “Cbam: Convolutional block attention module”. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018, pp. 3–19.
- [66] Changqian Yu et al. “Learning a discriminative feature network for semantic segmentation”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 1857–1866.
- [67] Jin-Gang Yu et al. “Exemplar-Based Recursive Instance Segmentation With Application to Plant Image Analysis”. In: *IEEE Transactions on Image Processing* 29 (2019), pp. 389–404.
- [68] Ze-Huan Yuan, Tong Lu, and Yirui Wu. “Deep-dense Conditional Random Fields for Object Co-segmentation.” In: *IJCAI*. 2017, pp. 3371–3377.
- [69] Matthew D Zeiler and Rob Fergus. “Visualizing and understanding convolutional networks”. In: *European conference on computer vision*. Springer. 2014, pp. 818–833.
- [70] Chi Zhang et al. “CANet: Class-Agnostic Segmentation Networks with Iterative Refinement and Attentive Few-Shot Learning”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2019, pp. 5217–5226.
- [71] Chi Zhang et al. “DeepEMD: Few-Shot Image Classification with Differentiable Earth Mover’s Distance and Structured Classifiers”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, pp. 12203–12213.
- [72] Chi Zhang et al. “Pyramid graph networks with connection attentions for region-based one-shot semantic segmentation”. In: *Proceedings of the IEEE International Conference on Computer Vision*. 2019, pp. 9587–9595.
- [73] Dingwen Zhang, Deyu Meng, and Junwei Han. “Co-saliency detection via a self-paced multiple-instance learning framework”. In: *IEEE transactions on pattern analysis and machine intelligence* 39.5 (2016), pp. 865–878.
- [74] Dingwen Zhang et al. “Cosaliency detection based on intrasaliency prior transfer and deep intersaliency mining”. In: *IEEE transactions on neural networks and learning systems* 27.6 (2015), pp. 1163–1176.
- [75] Dingwen Zhang et al. “Detection of co-salient objects by looking deep and wide”. In: *International Journal of Computer Vision* 120.2 (2016), pp. 215–232.
- [76] Hui Zhang et al. “Mask SSD: An Effective Single-Stage Approach to Object Instance Segmentation”. In: *IEEE Transactions on Image Processing* 29.1 (2019), pp. 2078–2093.
- [77] Kaihua Zhang et al. “Deep object co-segmentation via spatial-semantic network modulation”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. 07. 2020, pp. 12813–12820.

- [78] Qiang Zhang et al. “Revisiting Feature Fusion for RGB-T Salient Object Detection”. In: *IEEE Transactions on Circuits and Systems for Video Technology* (2020).
- [79] Qiang Zhang et al. “RGB-T salient object detection via fusing multi-level CNN features”. In: *IEEE Transactions on Image Processing* 29 (2019), pp. 3321–3335.
- [80] Xiaolin Zhang et al. “Adversarial complementary learning for weakly supervised object localization”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 1325–1334.
- [81] Xiaolin Zhang et al. “Sg-one: Similarity guidance network for one-shot semantic segmentation”. In: *IEEE Transactions on Cybernetics* 50.9 (2020), pp. 3855–3865.
- [82] Hengshuang Zhao et al. “Pyramid scene parsing network”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 2881–2890.
- [83] Bolei Zhou et al. “Learning deep features for discriminative localization”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 2921–2929.