

UNIVERSITY OF TECHNOLOGY SYDNEY
Faculty of Engineering and Information Technology

**LEARNING FOR OBJECT LOCALIZATION WITH
IMPERFECT DATA**

by

Xiaolin Zhang

A THESIS SUBMITTED
IN FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE

Doctor of Philosophy

Sydney, Australia

2021

CERTIFICATE OF ORIGINAL AUTHORSHIP

I, Xiaolin Zhang, declare that this thesis, is submitted in fulfilment of the requirements for the award of Doctor of Philosophy, in the School of Computer Science at the University of Technology Sydney.

This thesis is wholly my own work unless otherwise reference or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

This document has not been submitted for qualifications at any other academic institution.

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

Production Note:

Signature: Signature removed
prior to publication.

Date: 31 May 2021

Acknowledgements

I had a remarkable time at UTS pursuing my Ph.D. degree. My research journey have been truly amazing thanks to all the great people who have helped and supported me. I would like to express my sincere thanks to them all.

First and foremost, I would thank my supervisor, Prof. Yi Yang, for patient guidance and kind encouragement. He has given me a lot of suggestions for my research and future career. It is really the luckiest thing to have had him as my supervisor.

Also, I want to thank my co-supervisor, Dr. Yunchao Wei. He guided me through the realm of deep learning. His patience and kindness give me great support to complete every paper and finally finish my Ph.D. research work.

Then, I would like to thank the late Thomas S. Huang. I was so lucky to have met such a honorable scientist and have him as my advisor when I visited the University of Illinois Urbana-Champaign. His wisdom, kindness, and admirable faith in love really moved me.

And, I would like to thank my colleagues and friends for the help and support. I am delighted to have spent four great years in Sydney with all these lovely people.

I appreciate the financial support from the CSC-UTS Program.

Finally, I would like to express my deepest thanks to my parents, sister, and girlfriend, for their trust and love.

Xiaolin Zhang

November 2020 at UTS.

List of Publications

Conference Papers

- C-1. **X. Zhang**, Y. Wei, J. Feng, Y. Yang and T. Huang, “Adversarial Complementary Learning for Weakly Supervised Object Localization,” *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2018.
- C-2. **X. Zhang**, Y. Wei, G. Kang, Y. Yang and T. Huang, “Self-produced guidance for weakly-supervised object localization,” *Proceedings of the European Conference on Computer Vision*, 2018.
- C-3. **X. Zhang**, Y. Wei and Y. Yang, “Inter-Image Communication for Weakly Supervised Localization,” *Proceedings of the European Conference on Computer Vision*, 2020.

Journal Papers

- J-1. **X. Zhang**, Y. Wei, Y. Yang and T. Huang, “SG-One: Similarity Guidance Network for One-Shot Semantic Segmentation,” *IEEE Transactions on Cybernetics*, vol. 50, 2020.
- J-1. **X. Zhang**, Y. Wei, Z. Li, C. Yan and Y. Yang, “Rich Embedding Features for One-Shot Semantic Segmentation,” *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

Submitted Papers

- J-1. **X. Zhang**, Y. Wei, Y. Yang and F. Wu, “Rethinking Localization Map: Towards Accurate Object Perception with Self-Enhancement Maps,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Major revision, 2021.

Contents

Certificate	ii
Acknowledgments	iii
List of Publications	iv
List of Figures	ix
Abbreviation	xv
Abstract	xvi
1 Introduction	1
1.1 Background	1
1.1.1 Object Localization	1
1.1.2 Semantic Segmentation	3
1.1.3 Imperfect Data	4
1.2 Research Problem	6
1.2.1 Weakly supervised object localization	6
1.2.2 Few-Shot Semantic Segmentation	8
1.3 Thesis Structure	11
2 Literature Review and Related Works	13
2.1 Weakly Supervised Learning	13
2.1.1 Weakly Supervised Object Localization	13
2.1.2 Weakly Supervised Object Detection	15

2.1.3	Weakly Supervised Semantic Segmentation	16
2.2	Few-Shot Learning	18
2.2.1	Few-Shot Semantic Segmentation	18
2.2.2	Few-Shot Classification	19
2.2.3	One-Shot Video Segmentation	20
3	Adversarial Complementary Learning for Weakly Supervised Object Localization	22
3.1	Introduction	22
3.2	Adversarial Complementary Learning	24
3.2.1	Revisiting CAM	24
3.2.2	The Proposed ACoL	27
3.3	Experiments	29
3.3.1	Experiment Setup	29
3.3.2	Comparisons with the State-of-the-arts	30
3.3.3	Ablation Study	33
3.4	Summary	35
4	Self-Produced Guidance for Weakly Supervised Object Local- ization	36
4.1	Introduction	36
4.2	Self-Produced Guidance	37
4.2.1	Network Overview	37
4.2.2	Self-Produced Guidance Learning	39
4.3	Experiments	43
4.3.1	Experiment Setup	43

4.3.2	Comparison with the State-of-the-arts	43
4.3.3	Ablation Study	48
4.4	Summary	50
5	Inter-Image Communication for Weakly Supervised Localization	51
5.1	Introduction	51
5.2	Methodology	54
5.2.1	Object Seed Vectors	55
5.2.2	Stochastic Consistency	56
5.2.3	Global Consistency	57
5.3	Experiments	59
5.3.1	Experiment Setup	59
5.3.2	Comparison with the State-of-the-arts	60
5.3.3	Ablation Study	64
5.4	Summary	68
6	SG-One: Similarity Guidance Network for One-Shot Semantic Segmentation	70
6.1	Introduction	70
6.2	Methodology	73
6.2.1	Problem Definition	73
6.2.2	Proposed Model	74
6.2.3	Similarity Guidance Method	75
6.3	Experiments	77
6.3.1	Dataset and Metric	77

6.3.2	Implementation Details	78
6.3.3	Comparison	79
6.3.4	Multi-Class Segmentation	84
6.3.5	Ablation Study	85
6.3.6	Relationship with Video Object Segmentation	88
6.4	Summary	91
7	Rich Embedding Features for Few-Shot Semantic Segmentation	93
7.1	Introduction	93
7.2	Methodology	95
7.2.1	Problem Definition	95
7.2.2	The Proposed Method	97
7.2.3	The Network Structure	101
7.3	Experiments	104
7.3.1	Implementation Details	104
7.3.2	Comparison on PASCAL-5 ⁱ Lg	105
7.3.3	Comparison on COCO-20 ⁱ Lg	108
7.3.4	Ablation Study	109
7.3.5	Discussion	111
7.4	Summary	112
8	Conclusions and Future Work	115
8.1	Conclusions	115
8.2	Future work	117

List of Figures

1.1	Example of the object detection task.	2
1.2	Example of the semantic segmentation task.	3
1.3	Several techniques where imperfect data are involved. This thesis focuses on weakly supervised learning and few-shot learning.	4
1.4	This thesis focuses on the Weakly Supervised Object Localization (WSOL) and Few-Shot Semantic Segmentation (FSSS) problem. In WSOL, three approaches are proposed, <i>i.e.</i> , ACoL, SPG and I ² C. In FSSS, two methods are proposed, <i>i.e.</i> , SG-One and REF.	6
1.5	The framework of the WSOL methods.	7
1.6	An example framework, <i>i.e.</i> , SG-One (Zhang et al., 2020c) of FSSS. . . .	8
3.1	Comparison of methods for generating localization maps. Our method can produce the same-quality maps as CAM (Zhou et al., 2016) but in a more convenient way.	25
3.2	Overview of the proposed ACoL approach. The input images are processed by Backbone to extract mid-level feature maps, which are then fed into two parallel-classifiers for discovering complementary object regions. Each classifier consists of several convolutional layers followed by a global average pooling (GAP) layer and a softmax layer. Different from Classifier A, the input feature maps of Classifier B are erased with the guidance of the object localization maps from Classifier A. Finally, the object maps from the two classifiers are fused for localization.	26

4.1	Learning process of Self-produced guidance. Given an input image, we first generate corresponding attention map through a classification network. Then the attention map is roughly split, following the rule that the region with high confidence should be the object, whereas that with low confidence should be background. The regions with medium confidence remain undefined. All these three regions constitute the seed. Self-produced guidance is defined as the multi-stage pixel-level object mask supervised by the seed.	37
4.2	Overview of the proposed SPG approach. The input images are processed by Stem to extract mid-level feature maps, which are then fed into SPG-A for classification. Localization map is then generated from the feature maps from the last convolutional layer. Self-produced guidance maps are then calculated according to the map, then it is gradually learned using SPG-B. SPG-C utilizes the self-produced guidance map to train the classification network for learning pixel correlation. GAP refers to global average pooling.	38
4.3	Illustration of the attention maps and predicted bounding boxes of the SPG network on ILSVRC and CUB-200-2011. The predicted bounding boxes are in green and the ground-truth boxes are in red.	46
4.4	Output maps of the proposed SPG approach. The localization maps usually only highlight small region of the object. We extract the seeds of the self-produced guidance by segmenting the confident regions of the localization maps as foreground (white) and background (black), and ignore the left regions (grey). These seeds are applied as supervision to learn better self-produced guidance maps. Finally, the learned maps are leveraged to encourage the network to improve the quality of the localization maps.	47

5.1	(a) Convolutional operations preserve the relative pixel positions. Inconsistent response scores of different pixels on the class activation map are essentially caused by the inconsistent learned features. (b) The proposed Stochastic Consistency (SC) and Global Consistency (GC) are to align the object-related feature vectors.	52
5.2	The structure of the proposed approach. Given a pair of images (I_i^y, I_j^y) of the same category y , the localization maps (M_i^y, M_j^y) can be obtained by forwarding them through the classification network f_θ . Object seed vectors (V_i^y, V_j^y) are extracted from the high-level feature maps (F_i, F_j) according to the confident regions in the maps. Finally, the SC loss is employed on the object seed vectors. Also, the GC loss is employed on the averaged object feature a^y from a batch and the global class-specific center w^y . GAP refers to Global Average Pooling. AVG refers to the average operation.	55
5.3	Comparison of the predicted bounding boxes with ACoL (Zhang et al., 2018a). Our method obtains better localization maps and better bounding boxes. <i>The predicted boxes are in green and the ground-truth boxes are in red.</i>	62
5.4	(a): histogram of the number of object pixels with the threshold of 0.7 on the ILSVRC training set. (b): identified object regions (in red) according to the localization maps	67
5.5	Classification (<i>left</i>), localization (<i>middle</i>) and Gt-known Loc (<i>right</i>) error rates with the changes of hyper-parameters, <i>i.e.</i> , λ_1 , λ_2 and K	68

6.1	The network of the proposed SG-One approach. A query image and a labeled support image are input into the network. Guidance Branch is to generate the representative vector of the target object in the support image. Segmentation Branch is to predict the segmentation masks of the query image. We calculate the cosine distance between the vector and the intermediate features of the query image. The CosineSimilarity maps are then employed to guide the segmentation process. The <i>blue arrows</i> indicate data streams of support images, while the <i>black</i> are for query images. Stem is the <i>conv1</i> to <i>conv3</i> of VGG16. Interp refers to the bilinear interpolation operation. Conv is a convolutional block. Conv $k \times k$ is the convolutional filter with a kernel size of $k \times k$	72
6.2	Segmentation results on unseen classes with the guidance of support images. For the failure pairs, the ground-truth is on the <i>left</i> side while the predicted is on the <i>right</i> side.	80
6.3	Qualitative illustration of the one-shot and five-shot segmentation.	81
6.4	Similarity maps of different categories. With the reference to the support objects, the objects in the query images of the same categories will be highlighted, while the distracting objects and the background are depressed. The predicted mask can precisely segment the target objects under the guidance of the similarity maps.	81
6.5	Comparison between the few-shot image segmentation and few-shot video segmentation tasks. The object and background environment keep consistent in between video frames, while both objects and environment are greatly various in the image segmentation task.	89

- 7.1 An overview of our Rich Embedding Features (REF) approach. Given a query image from an unseen category, *e.g.* cat, its semantic mask is precisely segmented with the reference to only one annotated example of this category. The proposed REF module are employed to extract rich embeddings of the annotated support image. We then calculate the similarity maps by computing the feature distance between the embeddings and features of the query image. Similarity Maps are further applied to help the network segment the target object. 94
- 7.2 The overview of the proposed REF. (a) Given a pair of images including a query image and an annotated support image, they are first fed into the Stem module to obtain low-level features. Then, two branches, *i.e.*, Guidance Branch and Segmentation Branch, are applied to compute the guidance maps and the final masks. The input of Guidance Branch is both features of the query and support images. The corresponding output of the support image is fed into the proposed REF module to obtain the three kinds of embeddings. The final outputs in Guidance Branch are Similarity Maps obtained by computing the similarity distance between support embeddings and query features. The input of Segmentation Branch is the low-level query features and the high-level features from Guidance Branch. Similarity Maps can guide the segmentation process to focus on the targeted object regions by multiplying the maps with the acquired feature maps. We further propose to employ the depth-priority context module to enhance the segmentation masks. The *black arrows* indicate data streams of support images, while the *red* ones are for query images. (b) The proposed three kinds of rich embedding features which are applied to capture the diverse features of support objects. 96
- 7.3 The input feature maps are downsampled via convolution filters of stride 2 for several stages. The output feature maps of these filters are then upsampled and concatenated before putting into another convolution layer for adjusting dimensions. 99

7.4	Segmentation results on unseen classes with the guidance of support images. The proposed REF approach can successfully predict the masks for the target objects with the guidance of only one image.	103
7.5	Failure cases where the predicted masks of query images do not well match the ground-truth masks. There are generally two reasons for the failure predictions. One is that the appearance of the mismatched objects are too similar to the target objects, and another one is that too many noises and obstacles interfere the model from predicting accurate masks.	104
7.6	Segmentation result comparison of the embedding methods. REF can obtain more robust segmentation masks with the combination of the three proposed embedding methods.	109

Abbreviation

AP - Average Precision

bbox - Bounding Boxes

CAM - Class Activation Map

CNN - Convolutional Neural Network

DNN - Deep Neural Network

FCN - Fully Convolutional Network

FSL - Few-Shot Learning

FSSS - Few-Shot Semantic Segmentation

GAP - Global Average Pooling

IoU - Intersection-over-Union

PR - Precision Recall

WSL - Weakly Supervised Learning

WSOL - Weakly Supervised Object Localization

WSOD - Weakly Supervised Object Detection

WSSS - Weakly Supervised Semantic Segmentation

ABSTRACT

LEARNING FOR OBJECT LOCALIZATION WITH IMPERFECT DATA

by

Xiaolin Zhang

Deep learning has achieved countless remarkable successes in recent years. Learning deep neural networks usually needs tremendous well-labeled examples, which requires intensive investments. A feasible solution for reducing the budget is to learn from imperfect data, *e.g.*, noisy data, synthetic data, weak labels, and datasets with few annotated examples. This thesis dedicates to the weakly supervised learning and few-shot learning.

The first task is to address the challenging object localization problem using weak annotations as supervision. Objects in images are expected to be precisely located with only image-level labels, *i.e.*, category information. Specifically, convolutional networks can only find the most discriminative object regions leading to the unsatisfied predictions of bounding boxes. This thesis tries to solve this problem in three perspectives: 1) forcing the networks to mine more object areas by erasing the discovered object pixels; 2) learning pixel correlations within images under the supervision of self-produced object masks ; 3) communicating with different images to obtain more consistent features, and therefore, activating target object more accurately.

The second task is to predict the semantic masks of objects in a few-shot approach. Finding every pixel of target objects can also be considered as the most delicate localization problem. In the few-shot regime, only few annotated examples are available for an unseen class, and networks are required to locate the semantic category of each pixel with minimal information. This thesis will present two approaches to improve the quality of predicted object masks. Notably, a similarity-guided network is proposed to endow the segmentation process with rough position cues for locating the object pixels. To enhance the guidance process and improve the robustness, we further enrich the guidance embed-

dings and propose to employ multiple diverse support vectors to generate the similarity maps.

In addition, each of the proposed methods is comprehensively verified and analyzed by conducting various experiments.