Finding Four-Leaf Clovers: A Benchmark for Fine-Grained Object Localization

Laura Bravo* Alejandro Pardo* Gustavo Pérez* Pablo Arbeláez
Universidad de los Andes
{lm.bravo10, la.pardo2014, ga.perezs, pa.arbelaez}@uniandes.edu.co

Abstract

We present the Four-Leaf Clover (FLC) dataset, a new experimental framework for studying fine-grained object localization problems. We built the FLC dataset with the contribution of trained hobbyists, who were assigned the task of spotting four-leaf clovers on a fixed geographical extension over two clover seasons, one season for the train set and another for the test set. We then annotated each object instance for the tasks of object detection, semantic segmentation, instance segmentation, object parsing and semantic boundary detection. Our dataset is composed of more than 100,000 images, containing 2,151 carefully annotated clover instances of four, five or six leaves. The FLC dataset is extremely challenging and adapted to fine-grained object localization problems due to its small inter-class variance and its very large intra-class variation. We perform extensive experiments with state-of-the-art methods in order to establish strong baselines for each of the tasks.

1. Introduction

Fine-grained object recognition is a challenging open problem, in which the goal is to distinguish subordinate categories within entry-level categories. In other words, fine-grained recognition and, specifically, fine-grained localization is akin to looking for a needle in a haystack.

A large variety of applications require the study not only of fine-grained object categorization, but also of fine-grained object **localization**. Fine-grained localization is characterized by imbalanced problems with a small inter-class variance, and a large intra-class variation.

To the best of our knowledge, the computer vision community currently lacks a unified benchmark to study the different tasks associated with fine-grained localization. Table 1 shows a summary of the comparison between different publicly available datasets and the FLC dataset. On the one hand, datasets for fine-grained image categorization [9, 15, 19, 11, 20] have low inter-class variation, but the objects to be recognized are large and centered in the images, eliminating the object localization problem. On the other hand, datasets for object localization [3, 10, 17, 22, 1] exhibit high inter-class variation, making them inadequate for fine-grained problems.

In this paper, our main contribution is to present a comprehensive experimental framework, with highly detailed annotations, for studying fine-grained object localization tasks. We introduce the Four-Leaf Clover (FLC) Dataset, a novel benchmark for studying five different fine-grained localization tasks: object detection, semantic segmentation, instance segmentation, instance parsing, and semantic boundary detection. Additionally, we train and evaluate state-of-the-art methods for each task, particularly Mask R-CNN [7] for all the tasks, and also Convolutional Oriented Boundaries (COB) [12] and CASENet [21] for semantic boundary detection.

2. The Four-Leaf Clover Dataset

2.1. Dataset description

The FLC dataset was created as a benchmark to study fine-grained object localization. Table 2 shows statistics of the dataset, which is composed of more than 100,000 images, both positive, containing at least one four-leaf clover (for simplicity we refer to clovers with four or more leaves as four-leaf clovers), and negative, containing only three-leaf clovers. The authentic nature of each four-leaf clover was verified *in situ*, and negative images were captured after verifying that the patch did not contain any four-leaf clovers.

2.2. Fine-grained object localization tasks

The FLC dataset contains five fine-grained tasks with increasing level of detail. In this section, we describe each of these tasks and the evaluation methodology to assess them. The FLC annotation format is compatible with MS-COCO, in the interest of allowing a transparent use of the metrics and evaluation code proposed by this reference benchmark.

In the **object detection** task, the purpose is to determine

*Indicates equal contribution.

Table 1: Comparison of FLC to major visual recognition datasets. Club (♣) indicates that a dataset allows to study a recognition problem at a fine-grained level, triangle (△) indicates that the version of the problem is not fine-grained, and (×) indicates that a dataset does not allow to study a problem. The first six rows correspond to object recognition datasets that lack a fine-grained nature, while the next six rows present examples of fine-grained image classification datasets. (**CL:** Classification. **DE:** Detection. **SS:** Semantic Segmentation. **IS:** Instance segmentation. **BD:** Boundary Detection. **PS:** Parsing).

Dataset	CL	DE	SS	IS	BD	PS
Imagenet [2]	△	△	×	×	×	×
PASCAL [3, 14, 5]	△	△	△	△	△	×
MS-COCO [10]	△	△	×	△	×	×
DAVIS [17, 16]	△	△	×	△	×	×
ADE20K [22]	×	△	△	△	×	△
CityScapes [1]	×	△	△	△	×	×
iNaturalist [18]	♣	△	×	×	×	×
Cats & Dogs [15]	♣	△	△	×	×	×
CUB-200 [19]	♣	△	△	×	×	×
CompCars [20]	♣	△	×	×	×	×
VegFru [8]	♣	×	×	×	×	×
CDVCE [4]	♣	△	×	×	×	×
FLC	×	♣	♣	♣	♣	♣

Table 2: FLC dataset statistics. 4-leaf clover pixels and 4-leaf clover boundary pixels refer to the rate of the total of positive pixels over the total of pixels in the FLC dataset.

General statistics	Trainval set	Test set
Total positive images	1,000	500
Total negative images	51,637	51,670
Total images	52,637	52,170
4-leaf clover instances	1,530	747
4-leaf clover leaves	6420	3,128
4-leaf clover pixels	1.0511%	1.2431%
4-leaf clover boundary pxls.	0.0608%	0.0719%

the location of a four-leaf clover in an image. For the evaluation of an algorithm, we use the mean Average Precision (mAP) at a 0.5 overlap (IoU) as used in PASCAL VOC [3] (denoted as mAP@.5); and the averaged mAP with overlap of IoU ∈ [0.50: 0.05: 0.95] which is MS-COCO's [10] standard metric for detection.

In **semantic segmentation**, the goal is to classify all pixels corresponding to a set of classes, in our case, to classify all four-leaf clover pixels. For the evaluation of an algorithm, we use the mean intersection over union (mIoU) as used in PASCAL segmentation [3] and ADE20k [22].

In the **instance segmentation** task, the main goal is

Figure 1: Challenges of the FLC dataset. **(a)** Lighting and color variations. **(b)** Cast shadows. **(c)** Occlusion. **(d)** Leaf shape. **(e)** Leaf 3D orientation. **(f)** Different clover species.

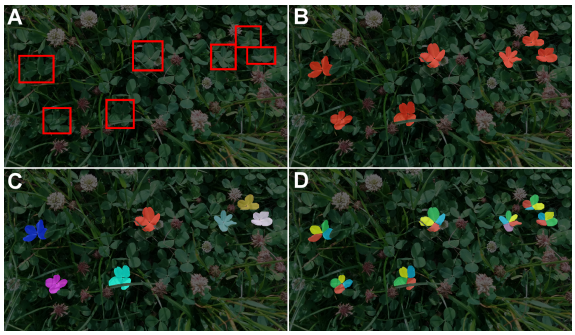

Figure 2: Annotation examples for four-leaf clovers in a sample image. **(A)** Detection. **(B)** Semantic segmentation. **(C)** Instance segmentation. **(D)** Parsing.

simultaneous detection and segmentation [6] of four-leaf clover instances. We evaluate the performance on our dataset using the same evaluation metrics of MS-COCO and PASCAL [6] for instance segmentation. MS-COCO uses Average precision (AP) at mask overlap (IoU) ∈ [0.50: 0.05: 0.95]. While PASCAL reports only AP at IoU at 0.5.

In **object parsing**, we are interested not only on the object instances but also on their parts. We want to detect each of the parts and be able to analyze how they interact at the whole object scale. We use the same evaluation metrics proposed in the Instance Segmentation task, setting as annotations the individual leaves of four-leaf clovers.

The aim of the **semantic boundary detection** task is to locate semantically defined boundaries, in particular only the boundaries of four-leaf clovers. The evaluation of the proposed methods for this task is a Precision-Recall (PR) curve, the area it encompasses (Average Precision (AP)) and the maximal F-measure (F1), introduced in the BSDS [13] and the SBD datasets [5].

2.3. Baselines

To quantitatively assess the need for a new dataset in which better fine-grained object localization methods can arise, and to set strong baselines for future reference, we

Table 3: Comparison of results using Mask R-CNN for the task of detection of four-leaf clovers

Dataset	Method	mAP@.5 (%)	mAP@[.5, .95] (%) all
FLC test (Positives)	Mask RCNN Res50+FPN	**56.4**	**35.6**
FLC test (Full)	Mask RCNN Res50+FPN	**6.20**	**4.10**

trained and evaluated state-of-the-art methods on every task available in the FLC dataset. Specifically, we used Mask-RCNN [7] for all the tasks and also COB [12] and CASENet [21] for semantic boundary detection. For every method, we report results of the complete test set and a subset of the test set composed of only positive images.

3. Experiments

3.1. Object Detection

Table 3 shows the results of evaluating Mask R-CNN for the detection task. We retrained starting from ImageNet weights, using Resnet-50 with Feature Pyramid Network (FPN) as the backbone.

3.2. Instance Segmentation

To address the problem of instance segmentation, we used the masks produced by Mask R-CNN [7]. We show the results in Table 4.

Table 4: Results of Mask R-CNN on the test set for the instance segmentation task. We show results only on the positive samples of the test set (FLC Test-Positives) and on the full test set (FLC Test).

Dataset	Method	mAP@[.5,.95](%)
FLC (Positives)	MaskR-CNN R50	**39.9**
FLC (Full)	MaskR-CNN R50	**4.7**

3.3. Semantic Segmentation

Table 5: Results using Mask R-CNN for the task of semantic segmentation of four-leaf clovers.

Dataset	Method	mIoU (%)
FLC (Positives)	MaskR-CNN R50	**32.71**
FLC (Full)	MaskR-CNN R50	**7.71**

We provide a baseline for the semantic segmentation task by using the outputs of our retrained Mask R-CNN models for instance segmentation as semantic segmentation masks. We report the results on the test set in Table 5.

3.4. Object Parsing

For this task, we used the same pipeline as the one used for instance segmentation with some modifications due to the increased difficulty of this task.

Table 6: Results of Mask R-CNN on the test set for the parsing task on only the positive samples of the test set (FLC Test-Positives) and on the full test set (FLC Test).

Dataset	Method	mAP@[.5,.95] (%)
FLC (Positives)	MaskR-CNN R50	**41.73**
FLC (Full)	MaskR-CNN R50	**1.3**

3.5. Semantic Boundary Detection

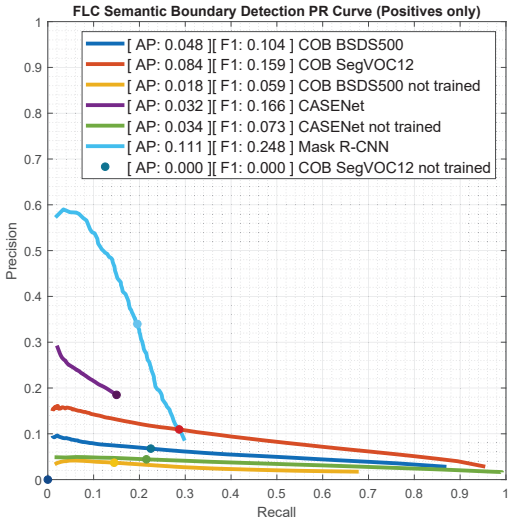

Figure 3: PR curves of 7 models for the Semantic Boundary detection task, evaluated on the positive images of the test set: untrained and retrained COB on the FLC (using the BSDS500 and SegVOC12 weights); untrained and retrained CASENet and Mask R-CNN.

4. Conclusions

We introduce the Four-Leaf Clover dataset, a new benchmark for studying fine-grained object localization problems. We propose an experimental framework for each task, testing state-of-the-art algorithms on each of them. Our results strongly indicate that the current best approaches from related recognition tasks are insufficient to solve fine-grained localization. Therefore, we hope that the availability of our dataset will spur the appearance of new ideas and methods for this largely unexplored and yet critical aspect of visual recognition.

References

[1] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The cityscapes dataset for semantic urban scene understanding. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1, 2

[2] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009. 2

[3] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html. 1, 2

[4] T. Gebru, J. Krause, Y. Wang, D. Chen, J. Deng, and L. Fei-Fei. Fine-grained car detection for visual census estimation. In *AAAI*, volume 2, page 6, 2017. 2

[5] B. Hariharan, P. Arbelaez, L. Bourdev, S. Maji, and J. Malik. Semantic contours from inverse detectors. In *The IEEE International Conference on Computer Vision (ICCV)*, 2011. 2

[6] B. Hariharan, P. Arbeláez, R. Girshick, and J. Malik. Simultaneous detection and segmentation. In *European Conference on Computer Vision (ECCV)*, 2014. 2

[7] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, pages 2980–2988. IEEE, 2017. 1, 3

[8] S. Hou, Y. Feng, and Z. Wang. Vegfru: A domain-specific dataset for fine-grained visual categorization. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. 2

[9] A. Khosla, N. Jayadevaprakash, B. Yao, and F.-F. Li. Novel dataset for fine-grained image categorization: Stanford dogs. In *Proc. CVPR Workshop on Fine-Grained Visual Categorization (FGVC)*, 2011. 1

[10] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: Common Objects in Context. In *European conference on computer vision (ECCV)*, 2014. 1, 2

[11] S. Maji, J. Kannala, E. Rahtu, M. Blaschko, and A. Vedaldi. Fine-grained visual classification of aircraft. Technical report, 2013. 1

[12] K.-K. Maninis, J. Pont-Tuset, P. Arbeláez, and L. Van Gool. Convolutional oriented boundaries. In *European Conference on Computer Vision (ECCV)*, 2016. 1, 3

[13] D. Martin, C. Fowlkes, D. Tal, and J. Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *The IEEE International Conference on Computer Vision (ICCV)*, 2001. 2

[14] R. Mottaghi, X. Chen, X. Liu, N.-G. Cho, S.-W. Lee, S. Fidler, R. Urtasun, and A. Yuille. The role of context for object detection and semantic segmentation in the wild. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. 2

[15] O. M. Parkhi, A. Vedaldi, A. Zisserman, and C. V. Jawahar. Cats and dogs. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012. 1, 2

[16] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross, and A. Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2

[17] J. Pont-Tuset, F. Perazzi, S. Caelles, P. Arbeláez, A. Sorkine-Hornung, and L. Van Gool. The 2017 davis challenge on video object segmentation. *arXiv preprint arXiv:1704.00675*, 2017. 1, 2

[18] G. Van Horn, O. Mac Aodha, Y. Song, A. Shepard, H. Adam, P. Perona, and S. Belongie. The iNaturalist Challenge 2017 Dataset. *arXiv preprint arXiv:1707.06642*, 2017. 2

[19] P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona. Caltech-UCSD Birds 200. Technical Report CNS-TR-2010-001, California Institute of Technology, 2010. 1, 2

[20] L. Yang, P. Luo, C. Change Loy, and X. Tang. A large-scale car dataset for fine-grained categorization and verification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 1, 2

[21] Z. Yu, C. Feng, M.-Y. Liu, and S. Ramalingam. Casenet: Deep category-aware semantic edge detection. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 1, 3

[22] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba. Scene parsing through ade20k dataset. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 1, 2